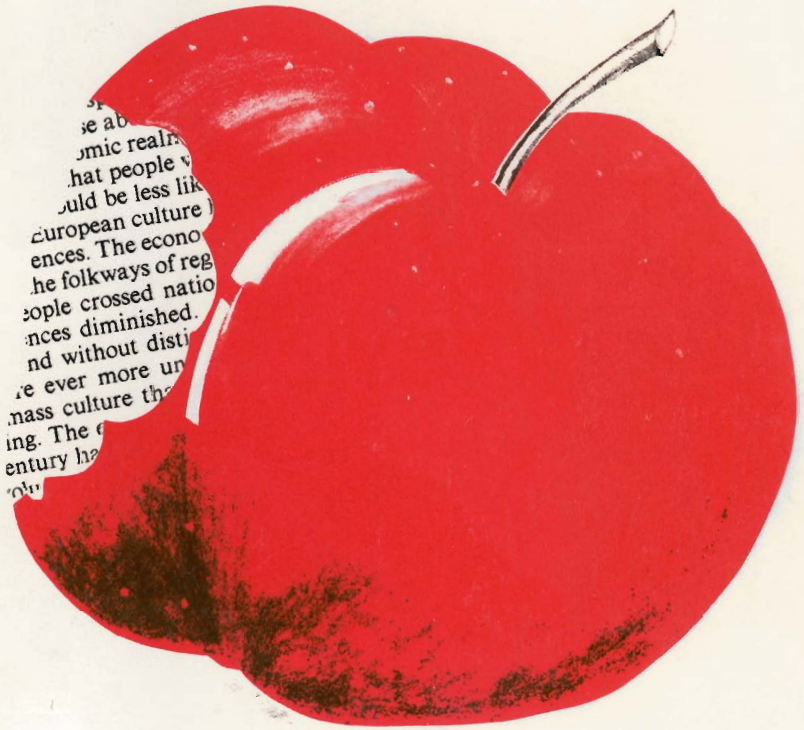
ICAME Journal

Computers in English Linguistics

No. 16

April 1992



International Computer Archive of Modern English

Norwegian Computing Centre for the Humanities

ICAME Journal

Computers in English Linguistics

No. 16

April 1992

International Computer Archive of Modern English

Norwegian Computing Centre for the Humanities

International Computer Archive of Modern English (ICAME)

ICAME is an international organization of linguists and information scientists working with English machine-readable texts. The aim of the organization is to collect and distribute information on English language material available for computer processing and on linguistic research completed or in progress on the material, to compile an archive of English text corpora in machine-readable form, and to make material available to research institutions.

The Norwegian Computing Centre for the Humanities in Bergen, Norway, acts as a distribution centre for computerized English-language corpora and corpus-related software. It publishes the *ICAME Journal* (previously *ICAME News*) and maintains an electronic information service (for details, see p. 140 in this issue). Conferences have been arranged since 1979.

ICAME ADVISORY BOARD

For information on ICAME, contact: Norwegian Computing Centre for the Humanities, Harald Hårfagresgt. 31, N-5007 Bergen, Norway.

Editors of the *ICAME Journal*:
Stig Johansson (Oslo)
Anna-Brita Stenström (Bergen)

Contents

Articles:

Merja Kytö and Matti Rissanen:
A language in transition: The Helsinki Corpus of English Texts 7

Geoffrey Leech and Roger Fallon:
Computer corpora – What do they tell us about culture? 29

Review articles:

Kari E. Haugland:
On the use of cleft and pseudo-cleft sentences in English (Peter C. Collins: *Cleft and pseudo-cleft constructions in English*) 51

Clive Souter:
The Nijmegen Linguistic Database program (Hans van Halteren and Theo van den Heuvel: *Linguistic exploitation of syntactic databases: The use of the Nijmegen Linguistic Database program*) 70

Rejoinder by Hans van Halteren 79

Reviews:

Karin Aijmer and Bengt Altenberg (eds.):
English corpus linguistics: Studies in honour of Jan Svartvik (Geoffrey Sampson) 81

Stig Johansson and Anna-Brita Stenström (eds.):
English computer corpora: Selected papers and research guide (W. Nelson Francis) 83

Merja Kytö:
Variation and diachrony, with Early American English in focus. Studies on CAN/MAY and SHALL/WILL (Udo Fries) 92

Ian Lancashire (ed.):
The humanities computing yearbook 1989-90 and *Research in humanities computing 1* (Stig Johansson) 96

Nelleke Oostdijk:
Corpus linguistics and the automatic analysis of English (Josef Schmied) 101

Felicitas Tesch:
Die Indefinitpronomina some *and* any *in autentischen englischen Sprachgebrauch und im Lehrwerken* (Magnus Ljung) 103

Gunnel Tottie:
Negation in English speech and writing: A study in variation (Leiv Egil Breivik) 105

Conference reports:

Lou Burnard:
Twelfth ICAME Conference, 1-12 May 1991 111

Stig Johansson:
The 'Using Corpora' Conference, Oxford 1991 113

Göran Kjellmer:
Nobel Symposium on Corpus Linguistics, Stockholm 4-8 August 1991 115

Anna-Brita Stenström:
Seminar on Corpus Studies and the Computer in English Language Research, Tampere 1991 117

Shorter notices:

Andrea Sand and Rainer Siemund:
LOB – 30 years on 119

Robert Davison:
Building a million-word computer science corpus in Hong Kong 123

Geoffrey Leech:
The Lancaster Parsed Corpus 124

Steve Crowdy:
The Longman/Lancaster English Language Corpus and the Longman Corpus of Learners' English 126

The British National Corpus 128

Paul Procter:
The Cambridge Language Survey 130

The Bank of English 133
The Georgetown University Catalogue of Projects in Electronic Text 134

Susan Hockey:
The Center for Electronic Texts in the Humanities 135

New Oxford Text Archive Catalogue 137
News on the Text Encoding Initiative (TEI) 138
The CHILDES CD-ROM 138
The ACL/DCI CD-ROM 139

Knut Hofland:
The ICAME CD-ROM 139

ICAME services:

Knut Hofland:
The ICAME network server 141

Texts available through ICAME 142

Conditions on the use of ICAME corpus material 145
Information for contributors 145

A language in transition: The Helsinki Corpus of English Texts[1]

Merja Kytö and Matti Rissanen
University of Helsinki

1. Overall dimensions of the Corpus

The diachronic part of the Helsinki Corpus of English Texts, henceforth the (Helsinki) Corpus, is a collection of early English texts now available in magnetic format from the Norwegian Computing Centre for the Humanities and the Oxford Text Archive.[2] The aim of this paper is to introduce the main characteristics of the Corpus and to discuss the rationale behind the text selection. Two examples are given to illustrate possible searches based on the material.

The main aim of the Corpus is to serve as a database for the study of the development of English morphology, syntax and vocabulary. The texts were selected in the spirit of sociohistorical variation analysis: they should give extensive evidence of varied types, modes and levels of linguistic expression. An attempt has been made to take the samples from good editions and to reproduce their spelling as accurately as possible. The extracts vary from 2,500 to some 20,000 words of continuous text; shorter texts are included *in toto*. No grammatical tagging can be offered at present, but parameter codings introducing each text give information on the author, type of text and discourse situation. The coding conventions and the lists of source texts are given in the *Manual*.[3]

2. Size and structure of the basic corpus

The diachronic part of the Helsinki Corpus consists of two entities, the basic corpus and the supplementary corpora. The basic corpus now available follows a systematic scheme of compilation and coding. The supplementary corpora are aimed at extending the temporal and geographical coverage of the Corpus, and should help develop and improve

the Corpus continuously. The main supplementary corpora in preparation at present, the corpus of Older Scots and the corpus of Early American English, will be integrated into the basic corpus in due course.

The basic corpus consists of c. 1.5 million words and is organized into three main sections: the Old English, Middle English and Early Modern (Southern) British English sections (see Table 1).[4] Old and Early Middle English texts are grouped into century-long subsections from mid-century to mid-century; later Middle English and Early Modern English periods are divided into subperiods of 70 or 80 years (see Table 1).

Table 1. The diachronic part of the Helsinki Corpus: size and period divisions (cf. Kytö 1991:2)

Subperiod		Words
OLD ENGLISH		
I	-850	2,190
II	850-950	92,050
III	950-1050	251,630
IV	1050-1150	67,380
		413,250
MIDDLE ENGLISH		
I	1150-1250	113,010
II	1250-1350	97,480
III	1350-1420	184,230
IV	1420-1500	213,850
		608,570
EMODE, BRITISH		
I	1500-1570	190,160
II	1570-1640	189,800
III	1640-1710	171,040
		551,000
		1,572,820

The size of the Corpus was originally modelled to follow the Brown and Lancaster-Oslo/Bergen corpora. However, the compilation work soon made it clear that a meaningful selection of texts would exceed the million-word framework. The present size of the Corpus is determined

by heuristic, practical, technical and financial factors. We are aware of the fact that a corpus of 1.5 million words is often too small for the purposes of diachronic research. This is particularly true of studies in which the distributions of variant forms are observed over successive periods of time. When dividing the Corpus, by way of illustration, into sections of a century, we arrive at subcorpora no larger than c. 150,000 words each. Moreover, the amount of linguistic evidence remains less substantial, because the material is broken down by various linguistic and extralinguistic parameters (e.g. by the type of text represented by more than twenty values, see Table 2). We welcome suggestions for additions to the current selection of texts and intend to produce an expanded and improved version of the Corpus within a period of five years or so.

In our experience the Helsinki Corpus yields a great deal of evidence and shows fairly consistent trends of development as regards a large number of linguistic phenomena. It should be clear to every user, however, that the Helsinki Corpus does not fully represent the English language of the past. We encourage the users to sharpen the picture given by the Corpus with further material that can be obtained from the texts themselves and the various editions, from other corpora, from printed concordances, and so forth.

The length of the individual samples included in the Corpus varies considerably. As certain texts and types of text were thought to be particularly important for diachronic study, they were sampled in greater length than others. Among these texts are, for instance, *Beowulf* and works by Chaucer. Prominence was also given to texts which have generally been thought to convey information on the forms and constructions typical of spoken language or the oral mode of expression (see below). Each text sample normally consists of at least two separate extracts.

3. A sample text

The parameter coding, which aims at giving shorthand information on the text, distinguishes the Helsinki Corpus from the previous multi-purpose corpora known to us. The codes are introduced in the COCOA format (cf. the Oxford Concordance Program) and can easily be modified for the purposes of other concordance programs. The following example, a handbook from the latter half of the 16th century, is taken from one of the source files:

<B CEHAND1B>
<Q E1 IS HANDO TURNER>
<N WINES>
<A TURNER WILLIAM>
<C E1>
<O 1500-1570>
<M X>
<K X>
<D ENGLISH>
<V PROSE>
<T HANDB OTHER>
<G X>
<F X>
<W WRITTEN>
<X MALE>
<Y 40-60>
<H PROF>
<U NON-PROF>
<E X>
<J X>
<I X>
<Z INSTR SEC>
<S SAMPLE X>

[^TURNER, WILLIAM.
TEXT: A NEW BOKE OF THE NATURES AND
PROPERTIES OF ALL WINES (1568).
A BOOK OF WINES.
EDS. S. V. LARKEY AND P. M. WAGNER (FACSIMILE).
NEW YORK: SCHOLARS' FACSIMILES & REPRINTS, 1941.
PP. B2R.1 - B8R.27 (SAMPLE 1)
PP. C6V.7 - D3V.19 (SAMPLE 2)
PP. D8R.2 - E1V.23 (SAMPLE 3)^]

<S SAMPLE 1>
<P B2R>
[}OF NEW AND OLDE WINE, AND OF IT
THAT IS OF A MEANE AGE THAT
IS NEYTHER TO BE CALLED
NEW NOR
OLDE.}]

There are twoo sortes of newe
Wine, one that is called (^Must^),
and that is but latelye made or
pressed out of the grapes, and is
swete in tast, troubled in color, and thick

in substaunce, and this sort is properlye
called in Latin (\Mustum\) . And another
sorte is called newe Wine, which hath
left his sweetnes & gotten clearenesse, but
yet it is not long since it was made. (^Galen^)
in his booke of making of medicines,
seemeth to call all Wine that is not fully
fiue yeares olde, newe wine, and it that
is past fiue yeares vntill it bee ten yeare
olde, wine of middle age, and it that is
aboue the age of ten yeares, olde wine,
and (^Dioscorides^) writing of the nature
of Wines in his fifte booke, calleth it
Wine of middle age, that is more than
seauen yeare olde, and (^Plinie^) writeth,

<P B2V>
not without an error of the scribe (as I
gesse) that (\Falerno media aetas incipit
ab anno decimo quinto\) . But (^Valeriola^)
a man otherwise wel learned, leauing
the authoritie of (^Galen^) , calleth it newe
Wine that keepeth still his Mustish and
swete taste, and as yet hath gotten no
sharpenesse, and he calleth that Wine of
middle age, that is no more sweete, but
is cleare, and sayth that he and his countrimen
take the most notable Wines of
Fraunce for olde Wines, before they bee
fullye one yeare olde. - - -

The top line in the code column describes the text in a nutshell and can be used as a reference code for the examples retrieved from the text by using OCP, WordCruncher, TACT, Corpus Manager (by Raymond Hickey) or other concordance programs (for the full lists of abbreviations used for reference codes, see the *Manual*). The above code values indicate, among other things, that the text is a handbook on a topic 'other' than 'astronomy' and 'medicine' (T), intended for layman users (U) and written by a male author (X) between 40 and 60 years of age (Y) from professional ranks (H). The parameter (C) defines the subperiod in which this text was written; for Middle and Early Modern English texts the more exact dates of composition or print are given in the *Manual*. A set of coding conventions are used to indicate headings [}...}], font other than the basic font (^...^), foreign language (\..\), our comments [^...^] and so forth (for details, see the *Manual*).

The parameter coding allows users to collect data in various ways. For instance, one can collect all the occurrences of the lexical item or structure in the entire Corpus or a part of it, and then work out how the relevant instances found are distributed as regards the various parameters (dialect, text type, author, etc.). Or, before collecting the examples, one can restrict the search to only those samples which fulfil a certain combination of values (e.g. religious treatises representing the East Midland dialect in 1250-1500).

It is in the nature of diachronic study that the further back in time we go, the harder it is to state solid facts about the status of a text. With Old and Middle English texts, in particular, lack of information has forced us to resort to the value X, by which we indicate that the parameter in question is either 'not applicable or irrelevant to this sample' or, 'not known' or that 'the information we have is too uncertain or inaccurate to be coded'. Revisions of many of the parameters will be necessary as our philological and linguistic knowledge of the texts increases. Further, despite our best efforts and two proofreadings, some erroneous codes and misprints could not be avoided. These will be corrected when the next version of the Corpus is released. All in all, the pilot studies completed so far have proved our coding to be useful in variationist analysis of corpus texts.[5]

4. Textual parameters and selection of texts

The representativeness of a diachronic multi-purpose corpus is largely a question of coverage. In our work we approached the issue from the following four angles:

(1) chronological coverage, to make the Corpus representative of all relevant periods and subperiods.

(2) generic coverage, to include samples representing a wide variety of genres or types of text.

(3) regional coverage, to take into consideration the regional varieties of the language.

(4) sociolinguistic coverage, to give room for texts produced by male and female authors representing different age groups, social backgrounds and levels of education.

The above criteria are listed roughly in order of importance when texts have been selected. This means that the date of the text has been of decisive importance. Next we have favoured a diverse selection of different types of texts, even though this practice may have caused

some imbalance in the way the Old and Middle English regional dialects are represented in the Corpus. Finally, with a few exceptions, we have given priority to prose texts and avoided verse and translated texts (see below).

(1) Chronological coverage

Among the main problems in the chronological coverage of a long-span diachronic corpus are the lack of the earliest (and sometimes even suitable later) texts, and the questions of how to divide the time into subperiods and how to deal with the differing dates of the hypothetical and extant manuscript versions of the texts.

Not surprisingly, a balanced and symmetrical chronological coverage is beyond our reach. The amount of text available from the 8th and 9th centuries, from the times of the Conquest, and from the 13th and early 14th centuries, is scanty, and these subperiods (OE1, OE2, OE4; ME1 and ME2, see Table 1 above) remain under-represented. We did not want to include excessively long samples from the few texts preserved to us from these periods; we believe that this would have endangered the comparability of the subsections to too great an extent.

As for defining the subperiods, a certain degree or arbitrariness cannot be avoided (for subperiod divisions, see Table 1 above). We have leaned heavily on sociohistorical factors and, when necessary, given up systematic and symmetrical solutions (for instance, taking the 1420s as a dividing line makes it easier to observe the emergence of standard English).

The difference between the dates of the original and manuscript version(s) of a text can extend from zero to centuries. As we believe that the users of the Corpus should be aware of this difference, we offer two code values when necessary. ME2/4, for instance, indicates that the original text probably dates from ME2 (1250-1350), while the manuscript included in the corpus was written in ME4 (1420-1500). When grouping the texts into subsections, the date of the manuscript has been followed.

(2) Type of text

No theoretically satisfactory model for classifying texts has so far been introduced for our purposes, which has been one of the reasons why we have resorted to heuristic rather than logical principles. To diminish circular definitions, our text type codings are based on extralinguistic criteria such as the subject matter and purpose of the text, the discourse situation and the relationship between the writer and the receiver. We

have, of course, profited greatly from earlier studies on register, formality, discourse types, etc.[6]

We strongly feel that some kind of text classification, albeit deficient in many respects, helps the users to approach the Corpus. It can also be used as a basis for future discussion of the typological features of the English texts of the past. Though we assume that the texts grouped under one and the same category most probably have some features in common, we do *not* claim that these texts would be uniform and homogeneous in terms of the distribution of linguistic or discourse features. On the contrary, the users of the Corpus should regard our grouping as one possibility among many. The text types distinguished in the Helsinki Corpus are given in Table 2 (cf. Kytö 1991:51-52).

Table 2. Text type

Old English	**Middle English**	**EMod English**
LAW	LAW	LAW
DOCUM	DOCUM	
HANDB ASTRONOMY HANDB MEDICINE	HANDB ASTRONOMY HANDB MEDICINE HANDB OTHER	HANDB OTHER
SCIENCE ASTRONOMY	SCIENCE MEDICINE	SCIENCE MEDICINE SCIENCE OTHER
		EDUC TREAT
PHILOSOPHY	PHILOSOPHY	PHILOSOPHY
HOMILY	HOMILY	
	SERMON	SERMON
RULE	RULE	
REL TREAT	REL TREAT	
PREFACE/EPIL	PREFACE/EPIL	
	PROC DEPOS	PROC TRIAL
HISTORY	HISTORY	HISTORY
GEOGRAPHY		

TRAVELOGUE	TRAVELOGUE	TRAVELOGUE
		DIARY PRIV
BIOGR LIFE SAINT	BIOGR LIFE SAINT	
		BIOGR AUTO BIOGR OTHER
FICTION	FICTION	FICTION
	ROMANCE	
	DRAMA MYST	
		DRAMA COMEDY
	LET PRIV LET NON-PRIV	LET PRIV LET NON-PRIV
BIBLE	BIBLE	BIBLE
X	X	

Some abbreviations:

BIOGR AUTO	= 'autobiography'
BIOGR LIFE SAINT	= 'biography: life of saint'
DIARY PRIV	= 'diary private'
DOCUM	= 'document'
DRAMA MYST	= 'mystery play'
EDUC TREAT	= 'educational treatise'
EPIL	= 'epilogue'
HANDB	= 'handbook'
LET PRIV	= 'private letter'
PROC DEPOS	= 'proceeding: deposition'
REL TREAT	= 'religious treatise'

It is apparent from Table 2 that we have aimed at a varied and diverse selection of the writings produced in all subperiods covered by the Corpus. We have included primarily non-literary texts but allowed room for literary genres as well. Given the effect of verse form on linguistic structure, the majority of the texts are in prose. However, a number of verse texts considered relevant have been included (e.g. OE poetry, *Layamon's Brut*, *The Ormulum*). Similarly, when no representative prose texts of an important genre have been available, verse has been selected (e.g. earlier ME romances, LME and early 16th-century drama).

As mentioned above, we have also tried to define the relationship of some texts to spoken language. This is mainly to overcome the major dilemma of variationist studies of the early periods, that is, the total lack of textual evidence of how people spoke in everyday life in the past. It is generally held that certain characteristics of the spoken idiom are reflected in text types such as sermons, trial records, drama, private letters, fiction and so forth. By comparing the results yielded by the 'non-speech-based' texts with those yielded by the 'speech-based' texts and 'scripts' (texts written to be spoken), we can see which expressions were likely to have been favoured and which avoided in speech.

Special attention has been paid to ensure diachronic and generic continuity over the three main periods represented (OE, ME, EModE). The eight text types that occur in all the three main sections of the Corpus are law, handbooks, science, philosophy, history, biography, fiction, and the Bible (with further subcategories). It has not, however, been possible to ensure full generic continuity over the individual subperiods within the three main periods.

The relatively large number of parameter values we use to code the (sub)periods and text types means that the break-down figures obtained for some linguistic items may remain low. One way of overcoming this problem is to resort to what we have experimentally defined as 'diachronic text prototypes'. Thus laws and documents have been given the code 'statutory'; handbooks, and some scientific, educational and philosophical treatises the code 'secular instruction'; homilies, rules, sermons and some religious treatises and prefaces the code 'religious instruction' and so forth. All in all, we have distinguished six prototypes, 'expository', 'imaginative narration' and 'non-imaginative narration' covering the rest. A number of text types remain outside the prototype categories; there can also be changes in the prototype values ascribed to certain texts, mainly owing to internal variation within a text.

Another way of trying to ensure the diachronic continuity in the Corpus can be seen in the two texts, i.e. the Bible and Boethius' *De Consolatione Philosophiae*, which have been sampled in several translations dating from different centuries. As for translations, though we have a code for defining the relationship of a text to a foreign original ('gloss' or 'translation') and give the source language(s) ('Latin', 'Latin/French', 'French', 'Dutch', 'other'), we have not collected translated texts in any systematic way.

(3) Regional dialect

Not surprisingly, the problems of providing the Corpus texts with reliable dialect codes are manifold. To begin with, the transmission history of a manuscript can be complicated, which makes it difficult to deal with scribal interference. Further, the extralinguistic evidence concerning early texts is often scarce or non-existent. Finally, the information given in reference works on the background of texts and authors may vary, ranging from vague and general statements to detailed studies of linguistic features. Consequently, the user of the Corpus must allow a certain degree of circularity in our dialect coding or turn to further sources to learn more about the background of the texts. The values used for the dialect coding are given in Table 3 (cf. Kytö 1991:50).

Table 3. Regional dialects

Old English	**Middle English**	**EMod English**
A/X	EML	ENGLISH
AM	EML/NL	
AM/X	EMO	
AN	WML	
K	WMO	
K/X	NL	
WS	NO	
WS/K	NO/EMO	
WS/A	SL	
WS/AM	SO	
WS/X	KL	
	KO	
	X	

Abbreviations: The slashes separate elements of mixed dialects. The final letter in Middle English dialect codings denotes the source of the definition: L = *Linguistic Atlas of Late Mediaeval English* (LALME); O = source other than LALME.

A	= 'Anglian'	EML, EMO	= 'East Midland'
AM	= 'Anglian Mercian'	WML, WMO	= 'West Midland'
AN	= 'Anglian Northumbrian'	NL, NO	= 'Northern'
K	= 'Kentish'	SL, SO	= 'Southern'
WS	= 'West-Saxon'	KL, KO	= 'Kentish'
		X	= 'unknown'

ENGLISH = 'Southern British standard'

All our samples of Old and Middle English have been given dialect or localization parameter values, while the post-1500 texts have been selected to represent the standard. The dialect codings of most earlier Middle English texts are based on the definitions found in the *Middle English Dictionary*; for later Middle English texts, the *Linguistic Atlas of Late Mediaeval English* has been consulted.[7] The codes given to many 15th-century texts are based on external evidence and merely indicate that a text represents some stage of development in the Southern standard.

(4) Sociolinguistic and discourse factors

While it is true that we know something about such Old English authors as King Alfred and Archbishop Wulfstan, this information is far too scarce and haphazard to allow any meaningful sociohistorical generalizions. This is why we give systematic information on the sex, age (in twenty-year age groups) and social rank ('high'/'professional'/'other' etc.) of our authors only from the Middle English period on. As regards correspondence, we also indicate the relationship ('intimate'/'distant') between the sender and the receiver; official letters are 'distant' by definition, while the letters exchanged between core family members are defined as 'intimate'. The relationship between the writer and the addressee can also be described as 'equal', or the writer can be seen to write to a person who is his or her social superior ('up') or inferior ('down').

To describe the intended audience of handbooks and scientific or educational treatises, we use the values 'professional' or 'non-professional'. We also define the relationship of a text to spoken language by using a three-level scale ('written'/'speech-based'/'script') and the possible interactivity of the discourse by distinguishing the typically 'interactive' texts (plays, correspondence, court trials) from 'non-interactive' texts. Finally, to define the elusive concept of formality, we have based our dichotomy 'formal' vs. 'informal' on the extra-linguistic factors involved in the discourse situation. Thus sermons, trial records and

official correspondence have been coded as 'formal', while private correspondence, comedy and light fiction have been coded as 'informal'.

5. Sample concordances

To conclude, we give some examples of possible searches. In Example 1, OCP has been used to retrieve the forms of the present-day WIT from the Early Modern English history writing, handbooks, and scientific and educational treatises (the relevant spellings were first checked in the *Oxford English Dictionary, Middle English Dictionary* and the WordCruncher version of the Corpus).

Example 1.

			wit 19
E3 NN HIST BURNETCHA	1,I,170	earance, a lively	wit, and a cheerful temper
E3 NN HIST BURNETCHA	1,I,169	uch levity in his	wit, and did not always ob
E2 EX EDUC BACON	20R	Bookes. For the	wit and minde of man, if i
E3 IS HANDO WALTON	214	he company with	(^wit^) and (^mirth^) , and
E3 NN HIST BURNETCHA	1,I,171	h he pretended to	wit and politics, he was n
E2 EX EDUC BACON	21R	aried trauaile of	wit, had ioyned varietie a
E1 NN HIST MORERIC	56	ior. For a proper	wit had she, & # could bot
E1 NN HIST MORERIC	40	er wit or trouth.	Wit if they were so dul, t
E2 EX SCIM CLOWES	9	is manyfolde: to	wit, inwardly and outwardl
E2 EX SCIM CLOWES	24	ough his singular	wit, long experience, and
E2 EX SCIM CLOWES	23	ation, that is to	wit: Not to leaue the Pati
E2 EX EDUC BACON	23R	n spent about the	wit of some one; whom many
E1 NN HIST MORERIC	46	after that is to	wit, on the friday the thi
E2 IS HANDO MARKHAM	72	ps of leather; to	wit one of them to his nee
E1 NN HIST MORERIC	40	, lacked # either	wit or trouth. Wit if they
E2 EX EDUC BACON	19V	nite agitation of	wit, spin out vnto vs thos
E1 IS/EX EDUC ASCH	279	il to goodnes, or	wit to learning, that coul
E2 EX EDUC BRINSLEY	14	t little ones: to	wit, to teach children how
E1 NN HIST MORERIC	78	so longed sore to	wit what they mente, gaue
			wits 12
E2 EX EDUC BACON	19V	harpe and stronge	wits, and aboundance of le
E2 EX EDUC BACON	23R	in the later many	wits and industries haue b
E2 EX EDUC BACON	23R	n the former many	wits and industries haue h
E2 EX EDUC BRINSLEY	12	charging of their	wits and memories. (^Phil
E3 IS EDUC HOOLE	10	h but amaze young	wits, and our English char
E2 EX EDUC BRINSLEY	11	in some pregnant	wits, and who are # indust
E2 EX EDUC BACON	19V	eading; but their	wits being shut vp in the
E2 IS HANDO GIFFORD	B3V	, or out of their	wits? (^Dan.^) I knowe tha
E3 IS EDUC HOOLE	1	titude of various	wits may be taught all tog

E3 IS EDUC HOOLE	215	part the choycest	wits, pickt out of other S
E3 IS EDUC HOOLE	6	pe do puzle young	wits to difference them, a
E2 EX EDUC BACON	22R	nd sober kinde of	wits; wherein the wisedome witt 1
E2 NN HIST HAYWARD	7	sexe; of # divine	witt, as well for depth of witte 11
E1 IS/EX EDUC ASCH	183	to sharpen a good	witte and encourage a will
E1 IS/EX EDUC ASCH	215	ill men: Men, of	witte and honestie, be oth
E1 IS/EX EDUC ASCH	185	wordes # without	witte. I wish to haue them
E1 IS/EX EDUC ASCH	279	his fellowes, in	witte, labor, and towardne
E1 IS/EX EDUC ELYOT	25	licate and tender	witte may be dulled or opp
E1 IS/EX EDUC ASCH	218	that haue neither	witte nor learning, to do
E1 IS/EX EDUC ASCH	217	uch, who haue not	witte of them selues, but
E1 IS/EX EDUC ELYOT	23	lackyng naturall	witte,) shall be apt to re
E1 IS/EX EDUC ASCH	186	th accompanie the	witte, there best vtteranc
E1 IS/EX EDUC ELYOT	27	efreshynge of his	witte, whan he hath tyme o
E1 NN HIST MORERIC	79	s euery manne may	witte, would neuer of like wittes 15
E1 EX SCIM VICARY	31	posed of the fiue	wittes, after the meaning
E1 EX SCIM VICARY	34	feebleness of the	wittes, and of al other me
E1 IS/EX EDUC ELYOT	29	excellent	wittes and vertuous occupa
E1 EX SCIM VICARY	32	surie of the fyue	wittes: And why he is an o
E1 IS/EX EDUC ASCH	185	red vp so in yong	wittes, as afterward they
E1 EX SCIM VICARY	31	e called the fyue	Wittes, as Hearing, Seeing
E1 IS/EX EDUC ELYOT	23	and in the tender	wittes be sparkes of volup
E2 IS HANDO GIFFORD	B4R	hich are in their	wittes for this worlde, wh
E1 IS/EX EDUC ASCH	280	o to norishe good	wittes in euery part of th
E1 NN HIST MORERIC	41	o # mistrust your	wittes, nor so suspicious
E1 EX SCIM VICARY	31	he fyue or common	wittes, or orgaynes, or in
E1 IS/EX EDUC ASCH	182	either dulled the	wittes, or taken awaye the
E1 EX SCIM VICARY	31	nd set the common	Wittes, otherwise called t
E1 EX SCIM VICARY	31	eth of the common	Wittes the fourme or shape
E1 EX SCIM VICARY	31	ued of the common	wittes withoutfoorth, repr wyttes 1
E1 IS/EX EDUC ELYOT	21	llectyue to noble	wyttes than to induce them

TOTAL WORDS READ	=	560946
TOTAL WORDS SELECTED	=	139231
TOTAL WORDS PICKED	=	59
TOTAL WORDS SAMPLED	=	59
TOTAL WORDS KEPT	=	59
TOTAL VOCABULARY	=	6

In Example 2 the combinations of (BE)CAUSE and THAT are retrieved using WordCruncher to illustrate the development of Present-day BE-

CAUSE. The relevant examples from the ME4 subperiod (1420-1500) are given in bold face italics:

Example 2.

Computer Book: **E:\HM4\HM4.BYB**
Reference List: **be-cause,be-cawse,be-kause,be-kawse,because, becavsse,becawis,bicause,bycause,by-cause, by-cawse,cause,cavs,cavse,cavsse,caws,cawse, kawse,that,thatt,+tat,+tatt,yt,ytt,y=t=,yat, yate**

Statut~ and orden=a=nc~, and considering **that** in div~s
partes #
of this yo=r= seid Realme ther be used mesures and weightes
som more large than the seid Standard and som lesse ***because***
#
that the very true mesure of the seid Standard is not to
all yo=r= true lieges verily knowen, at your owne p~pre cost
(M4 STA LAW STAT2 II,551:Heading)

tharchangell next co~myng. And **that** the seid Chief Officer
for the tyme beyng in ev~y suche Cite Towne or Borough have
for #
that cause a speciall Marke or Seale, to do marke
ev~y suche Weight and mesure so made to be reformed and
brought #
unto hym w=t=out fraude or delaye. And **that** he
(M4 STA LAW STAT2 II,552:Heading)

the seid suppliant+g (ben grevously vexed and labored
dailly) #
[\THE WORDS IN PARENTHESES CREASED\] and
so ben likly by longe tyme to endure. ***by cause that*** if +te
#
seid Thomas
Stamford perceyue **that** eny enquest woll not passe with his
(M4 XX DOC PET4 247:Heading)

any man
desire here after to Rauisshe Any woman) [\THE WORDS IN
#
PARENTHESES ARE OVER ERASURE\] and for **cause that** sche
wolle noghte assente (vnto hym sle and) [\THE WORDS IN
#
PARENTHESES RUBBED\] murdre her **that** any chartre
(M4 XX DOC PET4 262:Heading)

redresse is hadde vnto +te right intollerable hurt of all
the #
Comeyns of
this (Reame *by) **cause** +**tat*** many cloth makers +**tat** is to wete
#
men
weuvers fullers diers and women kempers Carders & spynners &
(M4 XX DOC PET4 268:Heading)

schuld be dysposyd **that** be born sundry days off the mone,
qwydyr to wurchyp or infortune; ye schal noght syngulerely
for trwth take **yt that yt** schuld be so; for euyl rwele may
cause that a man schal neuer come to wurchyp, thow he be
born to come to wurchyp; and off inffortune vndyr the lyke
forme.
But this ye may yeue for trwthe, as for a ryght dome,
(MX/4 IS HANDA METHD 155:Heading)

oyle of tartir, oyle of asshe, oyle of iunypre and soche
#
o+tere.
The **cause** why +**tat** +tai ben made oyles and the vertues
ben #
putte
in ham is twofolde: one **cause** is +**tat** it may bere +te vertue
(MX/4 EX SCIM CHAUL 580:Heading)

The **cause** why +**tat** +tai ben made oyles and the vertues
ben #
putte
in ham is twofolde: one **cause** is +**tat** it may bere +te vertue
#
+te more
depere. +Te secounde cause is +tat oyle schulde make +te
(MX/4 EX SCIM CHAUL 580:Heading)

in ham is twofolde: one **cause** is +**tat** it may bere +te vertue
#
+te more
depere. +Te secounde **cause** is +**tat** oyle schulde make +te
#
scharpenesse
|P581
(MX/4 EX SCIM CHAUL 580:Heading)

gouernour of all creatures, of whom all goodenesse comes;
and +tus +tou knalages is my+gthe. And se+t **+tat** he is
lorde and fadere, euery man owe+t hy[{m{] drede and
loue: drede ***by-cause +tat*** he is lorde, and worshippe
be-cause +tat he is fader.

These vij asskyngys of +te Pater Noster putte+t owte +te
vij dedely synnes and purchase+t +te vij +geftes of +te

(M3/4 IR SERM ROYAL 10:Heading)

and +tus +tou knalages is my+gthe. And se+t **+tat** he is
lorde and fadere, euery man owe+t hy[{m{] drede and
loue: drede **by-cause +tat** he is lorde, and worshippe
be-cause +tat he is fader.

These vij asskyngys of +te Pater Noster putte+t owte +te
vij dedely synnes and purchase+t +te vij +geftes of +te
#

(M3/4 IR SERM ROYAL 10:Heading)

and be +tis name we vndirstande +te ordr of Premonstracenses,
whech be-gan in Fraunce vndir a holy man +tei cleped
Norbertus,
+te +ger of our Lord a M and a hundred, and ***be-cause +tat*** I
mad
his lyf in Englisch to +te abbot of Derham **+tat** deyid last,
#

(M4 IR SERM CAPSERM 147:Heading)

langage grete mede for laboure; *+tis wil we applie to* ***+tat***
#
ordre
whech +tei clepe +te Freres of +te Crosse, for +tis **cause,**
for #
+tat
crosse on her breest schul make hem so to labour in +te weye

(M4 IR SERM CAPSERM 148:Heading)

#
sore
fore hor lyuelod: +tus +te lawe dyspensyth wyth apon hore
concyens. +Ten for ***bycause +tat*** Sonday ys no day of fastyng,
+terfor +ge schull begyn your fast at Aske-Wanysday, and
+tat #
day

(M3/4 IR SERM MIRK 82:Heading)

+tat
es at say, at +te pasch, als Haly Kyrke vses, when +tay ere clensede of syn thurghe penance, O payne of doynge owte of Haly Kyrke, bot if +tay forbere it by skillwyse **cause, +tat** awe
to be knawen to +tam **+tat** sall gyffe it; For he **+tat** tase it worthily, tase his saluacyone; and wha-so takes it
(M3/4 IR SERM GAYTR 8:Heading)

ende. the lesse he hath of wysdom. the worse he shal directe his dedes: but comynly erre. And very wyse men comynly ordre wel al theyr dedes & neuer erre And for this **cause. that** men wolde not erre from theyr ende. ye naturally entendyd. what zele & besynesse olde faders had to atteyne wisdom / wonder it is to rede. as at large declareth saynt Jerom in
(M4 IR SERM FITZJ B2V:Heading)

of substaunce, na partye of beynge, na it myght nought be
#
knawen
bot be the paynes **that** it is **cause** of. And this payne, it is sumthynge as to my syght, for a tyme: for it purges vs and makes vs to knawe oureselfe and aske mercy.
For the passion of oure lorde is comforth to vs agaynes
(M3/4 IR RELT JULNOR 60:Heading)

#
therfore
I make here an ende of this storie of Iason. whom diuerce menn blame ***because that*** he left & repudied Medea / but in this present boke ye may see the euydent causes / why he so dyd. Prayng my said lorde Prince taccapte & take **yt** in gree of me his indigne seruiteur. whom I beseche god
(M4 XX PREF CAXTON 36:Heading)

me semeth it is of grete nede / by **cause** I haue knowen it in my
yong age moche more welthy prosperous & rycher than it is at this day / And the **cause** is **that** ther is almost none / **that** #
entendeth
to the comyn wele but only euery man for his singuler
(M4 XX PREF CAXTON 77:Heading)

spyrytuel /
And as in my Jugement it is the beste book for to be taught
to yonge children in scole / & also to peple of euery age
it is ful conuenient yf it be wel vnderstanden / And by **cause**
I see **that** the children **that** ben borne within the sayd cyte
encreace / and prouffyte not lyke theyr faders and olders /
but
(M4 XX PREF CAXTON 77:Heading)

#
hand, +te
Frenschmen had gadered a gret nauy, with karikis and galeyes,
for to take Harflew. And for **+tat cause** +te kyng sent his
#
bro+tir,
Jon, duke of Bedford, with certeyn men of Ser Herry Percy,
(M4 NN HIST CAPCHR 247:Heading)

the #
laye
lP60
peple otherwhyle wexe wyse / the **cause that** thise clerkes
ben #
not
the wysest / is **that** they studye so moche in the connyng and
(M4 NI FICT REYNARD 60:Heading)

seyd I sopposyd **+tat** I xuld be here a fowrtennythe or iij
#
wekys. I pray +gou
+tat +te **caws** of my komyng away may ben kownsell tyl I speke
#
wyth +gou,
for +tei **+tat** lete me haue warnyng +ter-of wold not for no
(M4 XX CORP MPASTON 231:Heading)

+ter is no schepeherd but Hodgis sonys, for o+ter schepherd
#
dare non abyd
+ter ner com up-on +te comown ***be-kause +tat*** Wichyngham men
#
thretyn hem
to bete if +tei comen on here komon. And but if +gowr bestys
(M4 XX CORP MPASTON 232:Heading)

powre.
And Sire, my lady of Southfolke is halfindell dysplesyd
#
because that
my Cystere Barantyne is no better arayed, and leke wyse my
#
Cyster
(M4 XX CORP ESTONOR II,14:Heading)

ever affore, and **that** is a shrewde condiscion. Tell hym with
#
owte he
amend his condiscion **that** he *will* **cause** *strangers to*
advoide #
and come
no more there. I trust to you **that** he shall amend agaynest
(M4 XX CORP BETSON II,8:Heading)

#
conclevdyd
be Cortt **that** from Candyllmesse for+te no man shall sell but
ffor xxvj s. le li. I thynke **ytt** shall **cavsse** an stope. +Ge
#
most now
wrytt me yowr hadvysse how Y shall be demenyd: wher Y shall
(M4 XX CORP GCELY 97:Heading)

Notes

1. The project referred to as The Helsinki Corpus of English Texts: Diachronic and Dialectal was launched under the supervision of Professors Matti Rissanen (diachronic part) and Ossi Ihalainen (dialectal part) in 1984. This paper focuses on the diachronic part of the project. For references to reports and papers published on the work of both diachronic and dialectal parts, see Merja Kytö, *Manual to the diachronic part of the Helsinki Corpus of English Texts: Coding conventions and lists of source texts* (Helsinki: Department of English, University of Helsinki, 1991), p. 70, note 2. For the contributors to the work carried out on the diachronic part, see *idem*, pp. iv-vi. The Old English section profited especially from the work by Leena Kahlas-Tarkka, Matti Kilpiö, Ilkka Mönkkönen and Aune Österman. The Middle English section was largely compiled by Inkeri Blomstedt, Juha Hannula, Mailis Järviö, Leena Koskinen, Saara Nevanlinna, Tesma Outakoski, Päivi Pahta, Kirsti Peitsara and Irma

Taavitsainen and the section of Early Modern British English section by Terttu Nevalainen and Helena Raumolin-Brunberg. The Older Scots section is being compiled by Anneli Meurman-Solin and the Early American English section by Merja Kytö. Matti Rissanen supervised and coordinated the work, and Merja Kytö acted as the project secretary responsible for coordinating the team work and devising the database arrangements.

2. The various formats of the Helsinki Corpus offered are listed in the order forms available from the distributors.
3. The ideology and principles of compilation of the Corpus will be discussed in detail in *The Helsinki Corpus of English Texts: Introduction and pilot studies on the diachronic part*, eds. Matti Rissanen, Merja Kytö and Minna Palander, forthcoming.
4. The word counts exclude the portions of text coded as instances of 'foreign language', 'editor's comment' or 'our comment'. The Old English section of the Helsinki Corpus is based on the material taken from the text files of the Dictionary of Old English Project at the University of Toronto (Release 1, October 1982).
5. A list of the works based on the preliminary versions of the Corpus is given in Appendix 2 of the *Manual*, pp. 251-258.
6. E.g. Suzanne Romaine, *Socio-historical linguistics, its status and methodology* (Cambridge, London.: Cambridge University Press, 1982); Egon Werlich, *A text grammar of English* (Heidelberg: Quelle and Meyer, 1983 [1982]); M.A.K. Halliday and Ruqaiya Hasan, *Language, context, and text: Aspects of language in a social-semiotic perspective* (Geelong: Deaking University Press, 1985); Elizabeth Traugott and Suzanne Romaine, Some questions for the definition of 'style' in socio-historical linguistics, *Folia Linguistica Historica* VI, 1985: 7-39; James Milroy, Linguistic change, social network and speaker innovation, *Journal of Linguistics* 21, 1985: 339-384; and Douglas Biber, *Variation across speech and writing* (Cambridge, New York, etc.: Cambridge University Press, 1988).
7. *Middle English Dictionary*, eds. Hans Kurath, Sherman M. Kuhn et al. (Ann Arbor, Michigan: University of Michigan Press, 1954-); LALME = *A Linguistic Atlas of Late Mediaeval English*, eds. Angus McIntosh, M.L. Samuels and Michael Benskin (Aberdeen: Aberdeen University Press, 1986).

Computer corpora – What do they tell us about culture?

Geoffrey Leech and Roger Fallon
Lancaster University

1. The starting point for this study

The Brown Corpus and the LOB (Lancaster-Oslo/Bergen) Corpus are matching machine-readable text corpora of British and American English respectively.[1] Although these corpora were originally compiled for linguistic research, they may be regarded as a source of comparative information about varied social, political, and cultural aspects of the two most populous English-speaking countries. This paper reports a first systematic attempt to use them for this purpose.

Although we have called this the first 'systematic attempt', it is not entirely the first attempt, as this study was foreshadowed in Hofland and Johansson (1982), *Word frequencies in British American English*. This book largely consists of word frequency lists for the British (LOB) Corpus, but contains in one section (Ch. 8) a parallel alphabetical frequency list of both the Brown and LOB corpora. It also contains, in Section 3.5, an intriguing exemplification and discussion of some notable differences between the two corpora in terms of word frequency.

Since Ch. 8 of Hofland and Johansson's book is the chief starting point of this study, it is as well to begin with an extract from it, to show the way it is organized and the kind of information it gives:

word	*LOB*	*Brown*	*coeff*	*sig*
cottage	40	19	0.36	b
cotton	22	38	-0.27	c
couch	9	12	-0.14	
could	1614	1599	0.00	
couldn't	122	175	-0.18	b
council	343	103	0.54	a

council's	21	6	0.56	b
councillor	26	0	1.00	a
councillors	18	0	1.00	a
councils	35	6	0.71	a
counsel	11	17	-0.21	

The list shows, for each word listed, the absolute frequency of occurrence of the word in the LOB Corpus (on the left) and the Brown Corpus (on the right). A word is simply defined as an orthographic word-token. The list is restricted to more frequent words: those that occur at least 10 times and in at least 5 different text extracts in one corpus or the other. More interesting, however, are the two rightmost columns: first, there is a value varying between 1.00 and -1.00 representing the *coefficient* of *difference* between the two corpora, as calculated by the formula:

$$\frac{\text{Freq.}_{\text{LOB}} - \text{Freq.}_{\text{BROWN}}}{\text{Freq.}_{\text{LOB}} + \text{Freq.}_{\text{BROWN}}}$$

Figure 1: Difference coefficient formula

The extreme values yielded by this formula indicate that the word occurs either only in LOB (1.00), or only in Brown (-1.00). A positive value indicates a degree of 'overrepresentation' in LOB, and a negative value indicates a degree of 'overrepresentation' in Brown. If a word is equally well represented in both corpora, the formula yields a value of 0.00.

The very last column of the list contains (for some words only) a letter 'a', 'b', or 'c', indicating a chi-square value where the difference between the two frequency scores is significant at the following levels:

a = significance level .001
b = significance level .01
c = significance level .05

The significance level takes account of the absolute frequencies of words, but at the same time, the million words of each corpus is not a large enough sample to reflect reliably on the general differences between utilization of vocabulary in the U.K. and the U.S.A. As Hofland and Johansson warn us, it is probably safest to restrict attention to those items marked as showing significant difference of frequency, and we in general follow this practice in the following study. (Later we return to some reservations about the interpretation of these figures.)

The particular section of Hofland and Johansson which sparked off this study is Section 3.5 (pp. 32–40) in which they detail some of the

differences of vocabulary which emerge from the comparison of the corpora in Chapter 8, and which seem to invite explanation. For example, they note that the American corpus appears to be more extreme in its 'masculinity' than the British corpus: *he*, *boy* and *man* are more fully represented in Brown, whereas *she*, *girl* and *woman* are more fully represented in LOB. Such intriguing results (to which we return later) led us to attempt a more thorough and systematic study of vocabulary differences between the two corpora, to see what analytic classification could be proposed for the more salient differences, and (where appropriate) to hazard some explanations for them.

2. *Our goal: Using the corpora as evidence of cultural differences*

It is difficult to believe that there is an objective method for studying the numerous social, institutional, linguistic, and other factors which distinguish one culture from another. Yet, up to a point, the comparison of the Brown and LOB corpora does provide such an objectively-based method. The Brown and LOB corpora are a unique resource, both corresponding corpora being stratified random samples of written (printed) language selected from the same broad range of text categories, and according to almost exactly the same principles.[2] Significant differences of vocabulary between the two corpora are unlikely to be due to accidents of sampling, and therefore other explanations for these differences (e.g. cultural reasons) can reasonably be sought.

On the other hand, the method does have some real limitations. The most obvious of these are (a) the restriction of both corpora to written (printed) language only; (b) the restriction of the corpora to a particular historical period (the year 1961); (c) the restriction of the size of the two corpora to only c. a million words each. A million word corpus, although a large sample by most standards, is in fact relatively small for lexical research. For example, each corpus contains c. 50,000 word types, which is smaller even than the list of headwords found in the average desk dictionary, and much smaller than the total extent of the English lexicon. At the lower end of the frequency scale, in particular, the lexical differences between Brown and LOB do not give reliable guidance on American and British use of the language, since the relative frequency or infrequency of words may be due to accidents of sampling. But this skewing can be discounted at the higher end of the frequency scale, or where the comparison is based on whole groups of words, identified by semantic or other criteria.[3]

At this point, it will be useful to present a list of the contents of both corpora. The aim of the compilers was to cover a broad and fairly representative range of written language, sampled from 15 text categories (or *genres*).

Table 1: The basic composition of the Brown and LOB corpora

	Number of texts in each category	
	American	British
A Press: reportage	44	44
B Press: editorial	27	27
C Press: reviews	17	17
D Religion	17	17
E Skills, trades, and hobbies	36	38
F Popular lore	48	44
G Belles lettres, biography, essays	75	77
H Miscellaneous (government documents, etc.)	30	30
J Learned and scientific writings	80	80
K General fiction	29	29
L Mystery and detective fiction	24	24
M Science fiction	6	6
N Adventure and western fiction	29	29
P Romance and love story	29	29
R Humour	9	9
Total:	500	500

The 500 text samples consist of c. 2,000 words each.[4] Once the genre categories had been decided for the Brown Corpus, random sampling of bibliographies determined which text samples were included in each category. The LOB Corpus was compiled after the Brown Corpus, and reasonable steps were taken to ensure the contents of LOB corresponded to those of Brown as closely as possible.[5] However, minor discrepancies between the corpora do exist, and must be taken into account in judging the validity of comparisons.[6]

All in all, our confidence in the comparability of the Brown and LOB corpora remains fairly strong. One piece of evidence which supports this confidence is a very close match between the 50 most frequent items in both corpora: 49 out of the 50 words are common to both lists (see Hofland and Johansson, p.19). Another type of evidence giving confidence in comparability will become clearer as we proceed: it will be seen that, in areas where certain differences between American and

British English are well-known, and where certain predictable cultural differences can be expected, the corpora do indeed show these differences. The most obvious cases are known differences of spelling and lexical choice:

color	0	141	-1.00	a	
colour	140	0	1.00	a	
gasoline	0	12	-1.00	a	
petrol	12	0	1.00	a	

Proper nouns associated with the two nations also show expected differences:

Chicago	4	98	-0.92	a	
London	89	491	0.69	a	
Kennedy	85	140	-0.24	a	(US President in 1961)
Macmillan	59	1	0.97	a	(UK Prime Minister in 1961)

These are unsurprising and unexciting: but the more we find predictable differences showing up clearly in the corpora, the more we are inclined to trust as genuine the differences which have a less obvious explanation.

3. Research methods and research tools

3.1 Stage One

Our first and most important research tool is the one already mentioned: the comparative frequency list of words in the Brown and LOB corpora (Hofland and Johansson Ch. 8), together with the indices of significance. Our first step was to work through the whole alphabetical list, confining our attention to items marked by indices of significance as being favoured in one corpus or another. As we worked through this list, it became obvious that the items concerned grouped themselves into categories. We therefore set up provisional categories as we went along, and these were sometimes modified, rejected, or merged with others, as more items were examined.

It is convenient to use the word '*contrast*' (rather than 'difference') for a difference between the two corpora marked by a chi-square significance rating. In general, the contrasts could be divided into linguistic and non-linguistic types.

By *linguistic contrasts*, we mean contrasts obviously explained by differences between American and British English as language varieties. They were either differences in spelling (e.g. *theater, theatre*) or diffe-

rences in the choice of lexical items with the same meaning or reference (e.g. *transportation, transport*). Most of these differences were matters of frequency, rather than of total absence from one corpus:

transportation	3	43	-0.87	a
transport	64	18	0.56	a
movie(s)	7	60	-0.79	a
film(s)	244	133	0.20	a

Non-linguistic contrasts, on the other hand, here means contrasts which could not be easily explained as matters of linguistic code or variety, but where one had to postulate a difference of subject-matter – a difference in what was being talked about, e.g.:

coffee	54	78	-0.18	c
tea	111	60	0.60	a

As is fairly obvious, however, these illustrations reveal a difficulty of distinguishing the linguistic and non-linguistic categories, and more generally in interpreting the frequency lists. *Film(s)* and *movie(s)* are not complete synonyms, because *film* has a number of meanings which are unconnected with the cinema. Similarly, tea refers not just to a beverage which the British favour, but also the late-afternoon snack which commonly punctuates the British day (or, at least, used to in 1961). Because of multiple grammatical categories and multiple meanings, it is sometimes impossible, looking at the comparative frequency lists, to judge to what extent a contrast is due to a particular meaning, and hence to judge (in some cases) whether the contrast is linguistic or non-linguistic.

Further, multiple meaning causes a major problem for the classification of non-linguistic contrasts. The fact that the comparative list gives only information about graphic forms means that it is impossible, where a form is ambiguous, to assign it directly to one category rather than another. For example, without further evidence, *film* cannot be assigned to the category of 'Mass media'.

On these grounds, the search through the comparative frequency list could be only the first stage of a two-stage process. Before describing the second stage, however, we present a diagram of the major categories in which we placed the contrasts we found at stage 1:

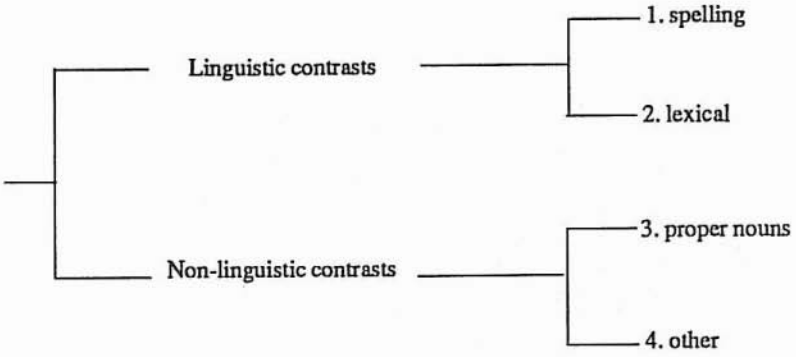


Figure 2: Main classification of contrasts

Of these four major categories, the first three could be dismissed as of little or no interest to this study. In order to explore socio-cultural differences, we naturally focused on the fourth category 'other'. This category we divided into sub-categories as follows, based on the domain with which the item is most associated:

1. Sport
2. Transport and travel
3. Administration and Politics
4. Social hierarchy
5. Military and violence
6. Law and crime
7. Business
8. Mass media
9. Science and technology
10. Education
11. Arts
12. Religion
13. Personal reference
14. Abstract concepts
15. *Ifs*, *buts*, and modality

(This list excludes some other domain categories which we arrived at during this search, but which we discarded as less easily identifiable and less important after further analysis.)

3.2 Stage Two

We turn now to the second and more detailed stage of the analysis (which in our case has been selective and incomplete). Here we made use of another research tool, the KWIC concordances of the Brown and LOB corpora. The following is an illustrative extract from the LOB KWIC Concordance (from Hofland and Johansson, p. 16). A complete

Figure 3: An extract from the LOB Concordance

L16	83	2	value of hypnotic treatment lies in the increased	suggestibility of the patient and also what we call
G55	54	1	listener absorbed. Sheer fatique might increase	suggestibility.) There was, in the Chinese manner, much
			WORDS = 2, G = 1, L = 1, SAMPLES = 2, G = 1, L = 1, CAT = 2	
BROWN			WORDS = 3, J = 3, SAMPLES = 1, J = 1, CAT = 1	
A17	54	9	public, which has shown itself infinitely	suggestible, knowing nothing between uncritical enthusiasm
F30	62	3	in his struggle for power with his brother, by	suggesting a deceitful way of taking possession of Leon A
G27	82	1	could well have dispensed with, and even humorously	suggesting a special service of intercession at OOSt Paul's
E10	170	8	The left arm could be making one of many gestures	suggesting excitement, and to link up with it the girl's
G69	158	7	Such magical, mysterious, awe-inspiring, divinity-	suggesting facts have included wholly outer phenomena like
B21	141	6	paper a description of Budapest in the early evening	suggesting it was a dark and depressing city. I can
G67	119	7	were capable of altering theoretical assumptions, by	suggesting new ones. Put in this way, Weber's procedure
C12	199	6	his own language, which seems hammered out, a medium	suggesting sheets of gold leaf. * <SEARCHER FOR ATLANTIS*>
A04	185	1	benefits paid during strikes and lock-outs.	Suggesting that a total æOT.U.C.á membership of eight million
D06	13	11	but to a child they are very real. Not that I am	suggesting that children should be molly-coddled*- they must
C13	128	1	he seems to want to dignify it*- oddly enough, by	suggesting that it is a poem of wit which, like Donne's is
J24	139	6	confirmed. It seems that the authors are correct in	suggesting that it is only rarely that average to bright
K06	9	1	probably thought a convulsive shudder. 'If you're	suggesting that Martin isn't old enough...' she said in a new
P12	97	4	bear to sit still. Even Bridget was no longer	suggesting that the girl had been caught out in an escapade.
B23	67	7	Clearly our two Democratic Socialists are	suggesting that the Labour Party should give up its heritage
C12	129	6	managed adroitly the humorous atrocity story	suggesting that the ministerial troops should castrate the
G47	55	5	is there. This irritation is sometimes relieved by	suggesting that the painting be viewed as a piece of wall
B10	15	7	in the accounts ØOMr. Aucott is not, I hope,	suggesting that the standards of honesty in British companies
D03	220	3	It carries the idea of reward or profit. Qoheleth is	suggesting that there is a gain from human experience. He has
F08	68	2	and asking him to accompany her, perhaps	suggesting that to make it entirely her evening, he allows
G73	101	10	seriously; it seemed to put responsibility on me for	suggesting what the difficulties might be and how they might
			WORDS = 20, A = 1, B = 3, C = 3, D = 2, E = 1, F = 2, G = 5, D = 2, E = 1, F = 2, G = 5, J = 1, K = 1, P = 1, CAT = 10	J = 1, K = 1, P = 1, SAMPLES = 19, A = 1, B = 3, C = 2, K = 1, L = 1, SAMPLES = 13, B = 1, D = 1, E = 1, F = 1, G = 2,
BROWN			WORDS = 13, B = 1, D = 1, E = 1, F = 1, G = 2, H = 2, J = 3, H = 2, J = 3, K = 1, L = 1, CAT = 9	
F40	151	11	from the plant is serious enough to merit almost any	suggestion aimed at controlling it, and this one, put forward
P07	139	12	for an early marriage. She was not adverse to the	suggestion, but he had to use a deal of pressure before she
J25	87	1	due to apparent size under limited conditions. This	suggestion could be tested experimentally by keeping one
J02	173	8	a warmer climate than today. This was the first	suggestion for an investigation into palaeoclimatology, a
F40	146	10	against them. Last week it was reported that the	suggestion had been put forward to use the coypu to combat
J38	196	2	be the case were the children not streamed. The	suggestion here is that since older children of a year group

KWIC Concordance of the LOB Corpus, compiled by Hofland and Johansson, is available both in machine-readable form and in microfiche form. (We used the microfiche version, although for a more extended study, the machine-readable version would prove the more powerful tool.)

For our purposes, there were two reasons for consulting the concordance: one was to check whether the frequency of the graphic form actually reflected the sense of the word we were interested in, for a particular category. For example, the higher incidence of *stress* and *pressure* in the Brown Corpus tempted us to place these items in a category of abstraction reflecting the supposed greater pace and intensity of life in the USA. Scrutiny of the concordances, however, caused us to reject this suggestion, since comparatively few of the occurrences of these words referred to the psychological domain, and collocations indicating physical stress or pressure were more common (e.g. *steam pressure, pressure on the trigger*).

The second reason for consulting the concordance was to check that the high frequency of an item was not due to any obvious skewing of its distribution in the corpus. Thus if a word with a frequency of 12 occurs 12 times in a single sample in the corpus, its incidence in the corpus obviously has much to do with an accident of sampling. (Such a word is *torsional* in the LOB Corpus.) But if a word with a frequency of 12 occurs (say) in 11 different text categories and 12 different samples, its occurrence in the corpus is clearly more representative. (Such a word is *emotion* in the LOB Corpus.) It is also possible to check how far a word's occurrences are concentrated in one or two text samples: for example, the high frequency of a particular word in the LOB Corpus may be largely due to the fact that in one text sample it occurs an extremely large number of times. We were able to note such cases, and eliminate the contrast containing them, if necessary, as due to sampling bias.

This information regarding the distribution of a word's occurrence throughout the genres can be directly checked from the frequency counts included in the concordance (see Figure 2). But we cannot tell from this how many occurrences are due to individual samples. For this purpose, it is necessary to scan all the examples in the concordance, and count the occurrences by hand. In addition, if we need to count the frequency of a word *in a particular sense*, as happens with polysemous words, this requires careful reading of the concordance examples, as well as counting. This requires an enormous amount of human labour, and in practice the task had to be simplified. To reduce the effect of

accidental clustering, we decided to discount any occurrences of words above the number of three in any given sample. The work that we began in the second stage has been carried out selectively, in cases where polysemy seemed especially likely to produce a misleading result. In other cases, we accepted the first-stage analysis as good enough for the exploratory purpose of this paper. As already stated, the task of comparing the domain categories in the two corpora is incomplete; but new research tools in the future will no doubt make the task much easier.[7]

4. Results: Contrasts between Brown and LOB in terms of domain

We will now illustrate each of the domain categories listed earlier, and where possible draw some tentative conclusions from the data presented. The general approach will be to begin with categories of a more *concrete* nature (in referring to physical realia), and to move to those which are *more abstract*.[8] This, in practice, means moving from relatively clear-cut and explicit categories (such as Sport) to categories which are relatively indeterminate and implicit (such as *Ifs*, *buts*, and modality). It also means progressing from the more obvious to the less obvious (and hence more intriguing, though tentative) findings.

In the lists which follow, the words cited are actually given with the frequencies in Hofland and Johansson Ch.8, even though the second stage analysis may have caused some modification to these. Words which were eliminated in Stage 2 as no longer yielding significant contrasts are omitted from these lists, which give only a selection of the more significant words in each domain category.

4.1 Sport

Under this heading, some of the obvious differences need scarcely be mentioned, such as the prevalence of *baseball* in Brown and of *cricket* and *rugby* in LOB. But more generally, more sporting terms prevail in the American corpus, conveying the impression that the American way of life has a more dominant interest in sporting activities. The following are among those sporting terms significantly more frequent in Brown: *athletic, ball, balls, game, games, golf, playing, pro, victor, victory, winning.*

4.2 Travel and transport

Again, the USA is a country where huge distances have to be covered by those moving from one place to another, and it seems natural that greater emphasis should be placed on travel and transport (or, to use the favoured American term, 'transportation'). Not only car travel, but travel by other means, is reflected in the list of terms significantly more frequent in Brown: they include *aircraft, auto, automobile, cars, highway(s), mileage, plane, river, trip(s), truck(s), vehicle(s), wagon* (all these are significance level 'a'). In return, the LOB Corpus can offer relatively few terms in this category, such as *canal, fares, airport, bus(es).*

4.3 Administration and politics

The next ten categories are somewhat less 'concrete': here are represented major sectors of public sociocultural life, such as government, the law, and religion. These are manifested in tangible and observable institutions, such as law courts and churches, but at heart they represent more abstract concepts such as democracy and justice.

Under the heading of Administration and politics, the Brown Corpus unsurprisingly favours *administration, congress,*[9] *governor, president, senate*, and *state*, referring importantly to US political institutions. These contrast with *council(s), ministry, monarchy, parliament* and *parliamentary*, which are favoured by the British corpus. Constitutional terms such as *congressional, legislative* and *resolution* are apparently favoured by the Americans, who also favour terms suggesting participatory and patriotic politics: *citizen(s), community/ies, leadership, nation(s), participation, public*. These contrast with the British preference for *authority* and *authoritative*.

Terms for political parties in the two corpora reflect, naturally enough, the party nomenclature of the two countries (*democrat* and *republican* versus *conservative, labour* and *tory*). It is less predictable, though not unexpected, that the American corpus favours *communist* and *communism* (communism being very much on the American mind in 1961) whereas the British one favours *socialist* and *socialism.*

4.4 Social hierarchy

It is again not surprising that LOB reflects the existence in the UK of the aristocratic hierarchy, with royalty at its apex. *King, queen, duke, duchess, earl, countess, royal*, and *empress* are all significantly more

frequent in LOB. To these may be added courtesy titles such as *sir, madam, lady* and *gentleman*, indicative of the respect accorded to social status.

In Brown, the two extremes of social scale (*president(s)* and *slave(s)*) receive prominence, but the great American populace between these has little or no differentiation. However, signs that American society may not be so egalitarian are glimpsed in the greater frequency of *status* and *elite*.

4.5 Military

The US corpus shows a particularly consistent and marked predominance in the use of military terms such as *armed, army, enemy, forces, missile(s)*, and *warfare*. (How overwhelming is the contrast between Brown and LOB in this respect can be seen in the Appendix.) In addition, the higher frequency of words concerned with firearms *(bullet(s), gun(s), rifle(s), shot)* will surprise no one aware that in the USA the 'gun' is a loaded concept. In contrast, one of the few terms in this area significantly more common in LOB is the negative term *disarmament*![10]

This American emphasis on military vocabulary is so consistent that it craves specific explanation. The year 1961 was the time of the Cuban missile crisis and shortly followed the building of the Berlin Wall. It was a high point of the Cold War, when the USA regarded itself as a policeman of the world perhaps even more than it has since.

4.6 Law and crime

The legal system is another area where the predominance of vocabulary in the American corpus is striking. This includes terminology specific to US law, such as *attorney, jurisdiction*, and *testimony*. But it also appears to indicate a greater general American involvement in legal matters, for example, in the commoner occurrence of *conviction, guilt, innocence, jury, justice, law(s), lawyer(s)*, and *trial(s)*. On the other hand, LOB favours only a few legal terms of relatively low frequency, such as *deposition, fines, imprisonment*, and *sentences*, apparently putting some emphasis on the penal function of the law.

The Brown Corpus also shows some concentration on crime and violence, showing a greater frequency of *killer, murderer, murders, policeman/men*, and *violence*. There is a link between this and the emphasis on firearms noted in the preceding section.

4.7 Business

Again, the American corpus favours business vocabulary. Examples of business terms significantly more frequent in Brown are: *bond(s), budget, business, corporation, costs, funds, loan, losses, management, manager, marketing, products, property, reserves, security, stock, stockholders.* The fact that the terminology of business (both commercial and financial) appears much more prevalent in Brown may be a sign of the greater sway of the business and of business ethic activities in the USA. (One wonders if the same prevalence would be evident today.)

It is tempting to go further, and to suggest that some terminological contrasts give evidence of attitudes underlying the UK's post-war commercial decline, at a time when the USA was unquestionably the world's dominant economic power. According to the corpus lists, Americans give importance to *input*, whereas the British do to *inCOME*! Often the business terms more frequent in LOB appear to concern financial benefits, rather than the effort which goes into production: e.g. *bonus, dividend, pension* and *remuneration.* Another sidelight may be found in the frequency of *export(s)*, marking the unusual importance of overseas trade to the British economy. And the predominance in Britain of a wage-earning and non-property-owning labour force is reflected in *earnings, wage(s), rents* and *tenants*, all significantly commoner in LOB.

4.8 Mass media

In the USA, 1961 was a period of all-powerful mass media, as is suggested by the prevalence in the Brown Corpus of *bulletin, editor, editorial(s), information, network(s), radio, reporters, journal, newspaper(s), advertising* and *journalism.*

4.9 Science and technology

The two corpora contrast very strikingly in terms of scientific terminology. But a more striking contrast still is in the area of technological vocabulary, where the Brown Corpus puts more focus, with terms such as *electronic(s), machine(s), plastic(s)* and *technology* itself. Perhaps linked to technology – specifically space technology – is an added preference, in the Brown Corpus, for cosmological terms, e.g. *astronomy, earth, mars, moon, planet(s), solar, space, sun, universe.*

4.10 Education

The differences between the two corpora in terminology of education accord in the main with well-known differences in the American and British education systems. The following are significantly more frequent in Brown: *campus, college, faculty, grade, graduate, graduation*. Moreover, the Brown Corpus appears to reflect a stronger awareness in American culture of the importance of education, and in particular of university education; not a surprising trend, considering the much greater proportion of the American population attending universities. General educational terms more prominent in Brown include: *class, classroom, scholars, scholarship, schools, studied/studies/study/studying*. Few educational terms are significantly more frequent in LOB, but one particular British emphasis seems to be on the titular results or awards of education, and the work one has to do to obtain them: *certificate, diploma, examination(s)*.

4.11 Arts

There is some evidence that the Americans have a greater predilection for the performing arts (*applause, concerts, dancer(s), drama, musician(s), orchestra*), whereas the British have a greater preference for books (*authors, book(s), read*). But there is much more work to be done on this category, and some of these differences may simply be due to accidents of sampling.

4.12 Religion

As Table 1 shows, both corpora contain religious texts as a separate text category. It is notable, however, that there is a greater preponderance of religious terms in the Brown Corpus as a whole, including words of key importance in the Christian religion: *Christ, Christian, churches, eternal, faith, God, Jesus*, and *religion* itself. In contrast, where the LOB Corpus favours religious terms, these tend to be terms such as *bishop, parish, vicar*, and the notoriously secularized *Christmas*, which signify outward and institutional aspects of religion. These tendencies correspond to one stereotypic view of the role of religion in the two nations: that the Americans are in general more deeply committed to the substance of religion, whereas the British are more concerned with the outward formalities.

4.13 Personal reference

We come finally to three categories which seem to lie on more general and abstract ground. Any explanatory comments here must be highly speculative.

As already noted, one of the most intriguing areas of comparison is in words, such as *man* and *woman*, which refer in a general way to people. Although there are some puzzling exceptions, the American corpus is more male-oriented than the British one. The following list of key gender-oriented terms is taken from Hofland and Johansson, p.38:[11]

he/him/his	17,603	19,412	-0.05
boy(s)	330	404	-0.10
man/men	1,789	2,113	-0.08
gentleman/men	61	49	0.11
she/her/hers	8,163	6,037	0.15
girl(s)	450	374	0.09
woman/women	486	468	0.02
lady/ies	184	122	0.20

As Hofland and Johansson point out, the male pronoun *he* with its oblique forms is over twice as frequent as the female pronoun *she* with its oblique forms in *both* corpora. But in Brown, the inequality of the sexes is even more salient: something which might suggest that the USA was in 1961 already ripe for the feminist movement which hit it in the later 1960s. (Again, it would be interesting to compare these figures with figures from the 1990s.)

Other male words particularly common in Brown confirm the stronger masculine bias: *boy(s)* and *man/men*. *Gentleman*, on the other hand (presumably for reasons of gentility rather than masculinity – see 4.4 above), goes against the trend and is overrepresented in the LOB Corpus, in company with *girl*, (marginally) *woman*, and (particularly) *lady*.

The pattern with family words such as *father* and *mother* is less dominated by gender: *father* and *mother* are both significantly more frequent in LOB, as are the family-related terms *marry* and *marriage*. One possible conclusion from this is that the stress on masculinity, particularly strong in Brown, is partly counteracted in LOB by an emphasis on family relationships.

4.14 Abstract concepts

We are now delving into more dangerously speculative regions. Looking

through the many abstract nouns which have significantly greater frequency in the American corpus,[12] one is struck by some groups as follows: (a) Brown favours grand abstractions, particularly those representing the ideals of a democratic society, e.g. *freedom, independence, justice, liberty.* (b) Another small group appears to support the view that the USA has been more attuned than the UK to the achievement ethic of enterprise culture; for example, Brown favours *effectiveness, efforts, planning,* and *project(s).* (c) The British corpus, on the other hand, seems to give preference to abstract categories representing emotions and attitudes, e.g. *disgust, doubt, goodwill, happiness, jealousy.*

4.15 Ifs, buts and modality

Again, in a speculative vein, we note that the concessive conjunctions *but, although,* and *though* are significantly more common in LOB, perhaps manifesting a British tendency to trim and to temporize – to see both sides of a question. The same lack of decisiveness may explain the higher frequency in LOB of the conditional conjunctions *if* and *unless.* Somewhat related to these are differences in modality and in the use of hedges. Generally words denoting possibility or uncertainty are more frequently used in the LOB Corpus, e.g. *impossible, improbable, perhaps, possible, possibly, unlikely. (Probable, probably* and *likely* are also more frequent in LOB, though not significantly so.) The exception here is the typically American adverb *maybe,* which is much more common in Brown. However, if we add together the frequencies of the synonyms *maybe* and *perhaps,* the frequency is still much higher in LOB, suggesting that *maybe* is an exception for dialectal reasons, and does not prove a counterexample to the general trend. The hedging adverbs *rather, quite* and *fairly* are also more strongly associated with the LOB Corpus, although this tendency is slightly counterbalanced by the greater frequency of *somewhat* in the Brown Corpus.

These features add up to a suggestion – no more than that – that the LOB Corpus shows conformity with one British stereotype, of the wishy-washy Briton who lacks firmness and decisiveness, seeing two sides to every question, and shades of grey instead of black and white.

5. Conclusion

Wrapping up the whole analysis of Section 4 in one wild generalization, we may propose a picture of US culture in 1961 – masculine to the point of machismo, militaristic, dynamic and actuated by high ideals, driven by technology, activity and enterprise – contrasting with one of

British culture as more given to temporizing and talking, to benefitting from wealth rather than creating it, and to family and emotional life, less actuated by matters of substance than by considerations of outward status. However much of a caricature, this is not an unconvincing portrayal for those of us who have lived with or in both cultures through recent decades.

However, to return to a note of caution. The method has been explored, but has not been fully tested. It is basically a technique of proceeding empirically, in a 'bottom-up' fashion, from what is indubitably there in the corpora, to what can only be inferred, or surmised to be the case. The technique claims to be moving towards a true picture of cultural contrasts, on the grounds that the evidence is in the corpora, and no other explanation can be found for it. Possible other evidence will show some of the conclusions tentatively arrived at in this paper to be false. This is in the nature of scientific progress!

There are many defects in the study, and some of them will no doubt be remedied in the near future. The use of a parallel frequency list based on the tagged Brown and LOB corpora would be an obvious step forward. In future, if such corpora are semantically tagged with word senses, progress will be even more substantial. In the future we look forward to parallel corpora including spoken as well as written material.[13] We also look forward to larger parallel corpora,[14] which will render obsolete the type of exercise in which we have been engaged, where one is basing the study on a limited range of vocabulary. Another clear enhancement of this kind of study would be a study which looks beyond frequencies of individual words in isolation, to frequency of collocations, or words in context.

Notes

1. See the Manuals of Information of these corpora: Francis and Kučera (1979) and Johansson et al (1978). The contents of the corpora are described in Table I on page 32.
2. In two rather minor aspects of sampling the two corpora do not correspond, because of different publishing practices: (a) the press genres (A-C) in LOB are sampled within national and regional categories, whereas this distinction does not exist in Brown; (b) in genre N 'Adventure and western fiction', the category of 'Western fiction', for obvious reasons, is well represented in Brown, but not in LOB.
3. Two minor factors may interfere with the comparability of the two

corpora in ways which have not yet been fully explored. (a) The LOB Corpus appears to contain somewhat more quoted material than the Brown Corpus, and (b) the sampling of the Brown Corpus was based on the catalogues of large libraries, whereas that of the LOB Corpus was based on national bibliographies.

4. In fact, the length of each text sample was in general slightly more than 2,000 words, since the end was taken to be the first sentence break on or after the 2,000th word.
5. See notes 2 and 3 above.
6. Already there exist lemmatized frequency lists (Francis and Kučera 1982, and Johansson and Hofland 1989) based on the grammatically-tagged versions of these corpora. These have been used in the present study, as a means of checking the frequency of meanings associated with particular word-classes (e.g. *general* as a noun, as contrasted with *general* as an adjective). On the basis of these lists, a parallel lemmatized frequency list for comparing the two corpora could be compiled, to avoid some of the labour of sorting out meanings now necessary with the list in Ch. 8 of Hofland and Johansson. (However, there are some differences between the grammatical tags used for tagging Brown and LOB, and so some difficulties would occur in the comparison.) On further corpus-based research tools for the future, see Section 5, and notes 12 and 13 below.
7. Discussions of culture make a distinction between 'material and non-material elements' (Young 1972), or between 'a notion of culture as observable phenomena' and 'a notion of culture as not observable: something which is internal but which can also be explicitly described' (Nemetz Robinson 1985). This study sees the study of culture as a synthesis of these two conceptions.
8. Note that Hofland and Johansson's list distinguishes only between (i) words which are always spelt with an initial capital (e.g. *Sam*) and hence appear with an initial capital in the list, and (ii) words which are sometimes or always spelt with an initial lower case letter, e.g. *young*. The latter words appear in the list with a lower case initial letter. The difficulty with this practice is that names identical in letter spelling to an ordinary word (e.g. *Young*, as compared with *young*) get merged with that word in the list. Hence the spelling *congress* occurs in the list, even though the word is usually spelt with a capital: *Congress*. This practice accounts for other words unexpectedy spelt with a capital in our lists: *arsenal, tory, mars*, etc.

9. The higher frequency of *arsenal* in the British corpus has less to do with any military installation than with a leading London football team!
10. See Kjellmer (1986) on gender bias in the Brown and LOB corpora.
11. The predominance of abstract concepts in the Brown Corpus as compared with the LOB Corpus is striking, and leads one to wonder whether there is not an underlying stylistic difference between the two corpora, with American writers showing a greater penchant towards nominalization or abstraction. Another possible explanation is that the sampling for the Brown Corpus, being based in part on the holdings of the Brown University library, led to a slight tendency toward the selection of more learned or scholarly material. This invites investigation.
12. The International Corpus of English (under the direction of Sidney Greenbaum and others) is planned to consist of at least 15 parallel corpora of English, collected not only from the USA and the UK, but from other major countries where English is a first or second language. These corpora will consist of both spoken and written material, and will invite further more broadly-based comparative studies of the kind we have undertaken here.
13. At present new English corpora and collections of machine-readable text are being compiled on a much larger scale than heretofore. The British National Corpus (a 100-million-word corpus of spoken and written British English – being compiled by a consortium of Oxford University Press, Longman, Chambers, Lancaster University, Oxford University, and the British Library) is one example. Others are the Bank of English (Collins and Birmingham University), and the ACL Data Collection Initiative (Liberman and others, University of Pennsylvania). In the 1990s we can look forward to lexical studies based for the first time on corpora of adequate size. One may hope that by the year 2000, it will be possible to make use of these corpora for cross-cultural studies on a much larger scale than is now possible on the limited basis of the Brown and LOB corpora.

References

Fallon, Roger. 1990. *The use of text corpora in the comparative study of American and British culture.* Unpublished M.A. dissertation, Department of Linguistics and Modern English Language, Lancaster University.

Francis, W. Nelson, and Henry Kučera. 1979. *Manual of information to accompany a Standard Sample of Present-day Edited American English, for use with digital computers.* Original ed. 1964, revised 1971, revised and augmented 1979. Providence R.I.: Department of Linguistics, Brown University.

Francis, W. Nelson, and Henry Kučera. 1984. *Frequency analysis of English usage: Lexicon and grammar.* Boston: Houghton Mifflin.

Hofland, Knut, and Stig Johansson. 1982. *Word frequencies in British and American English.* Bergen: Norwegian Computing Centre for the Humanities/London: Longman.

Johansson, Stig, and Knut Hofland. 1989. *Frequency analysis of English vocabulary and grammar.* Vols. 1-2. Oxford: Clarendon Press.

Johansson, Stig, Geoffrey Leech and Helen Goodluck. 1978. *Manual of information to accompany the Lancaster-Oslo/Bergen Corpus of British English, for use with digital computers.* Oslo: Department of English, Oslo University.

Kjellmer, Göran. 1986. 'The lesser man': Observations on the role of women in modern English writings'. In *Corpus linguistics II: New studies in the analysis and exploitation of computer corpora*, ed. by Jan Aarts and Willem Meijs. 163-76. Amsterdam: Rodopi.

Nemetz Robinson, G.L. 1985. *Crosscultural understanding: Processes and approaches for foreign language, English and bilingual educators.* Oxford: Pergamon Press.

Young, R.W. 1972. Culture. In *Language and cultural diversity in American education.* ed. by R.D. Abrahams and R.C. Troike. Englewood Cliffs, N.J.: Prentice-Hall.

Appendix: List of contrasts of vocabulary related to the Military domain

Frequency ratios are shown as x:y, x being the LOB frequency and y being the Brown frequency. (Significance levels are indicated by 'a' (.001), 'b' (.01), or 'c' (.05).

1. Words significantly more frequent in the Brown Corpus

aircraft	31:70	a	fighters	4:16	b
armed	22:60	a	force	168:229	b
ballistic	1:17	a	guerilla	1:12	b
bullets	4:21	a	lieutenant	12:29	b
cavalry	3:26	a	manned	2:12	b
code	13:40	a	missiles	11:32	b
column	24:71	a	missions	3:16	b
corps	6:109	a	patriot	0:10	b
enemy	38:88	a	rifles	6:23	b
fallout	1:31	a	signals	11:29	b
fire	125:187	a	slug	1:10	b
forces	84:175	a	troop	3:16	b
fort	11:55	a	veteran	11:27	b
guerillas	0:17	a	victor	8:23	b
gun	56:118	a	victory	32:61	b
guns	8:42	a	Viet	3:16	b
headquarters	28:65	a	armies	5:15	c
losses	11:46	a	arms	85:121	c
major	142:247	a	army	101:132	c
marine	12:55	a	assault	4:15	c
mercenaries	0:12	a	battery	6:18	c
military	133:212	a	bullet	12:28	c
militia	0:11	a	campaigns	5:17	c
missile	5:48	a	civilian	1:24	c
mobile	6:44	a	command	49:72	c
patrol	6:25	a	commands	5:15	c
plane	49:114	a	enlisted	3:11	c
rifle	20:63	a	fought	28:46	c
Sherman	0:29	a	infantry	6:16	c
shot	65:112	a	marching	5:15	c
squad	2:18	a	march	85:120	c
strategy	4:22	a	mission	49:78	c
submarine	7:27	a	peace	159:198	c

target	18:45	a	pentagon	3:13	c
veterans	1:16	a	pirates	3:12	c
volunteers	8:29	a	pistol	13:27	c
warfare	14:43	a	signal	40:63	c
battle	54:87	b	strategic	9:23	c
bombs	16:35	b	tactics	8:20	c
bombers	7:22	b	targets	9:22	c
combat	8:27	b	territorial	5:14	c
codes	4:17	b	war	396:464	c
destroy	25:43	b	weapon	24:42	c
division	64:107	b	Winchester	4:12	c

2. Words significantly more frequent in the LOB Corpus

medal	37:7	a	conquest	20:9	c
disarmament	27:11	a	rank	43:24	c
trench	15:2	b	tanks	35:18	c

Note: These lists illustrate one of the most extreme differences between the corpora: that of military terminology. The lists are those resulting from Stage 1 of the analysis; there is no guarantee that all the words are used entirely, or even predominantly, in a military sense.

Review articles

On the use of cleft and pseudo-cleft sentences in English

Kari E. Haugland
University of Bergen

Peter C. Collins. *Cleft and pseudo-cleft constructions in English.* London & New York: Routledge, 1991. (Theoretical Linguistics series.) ISBN 0–415–06328–0. 230 pp.

0. Introduction

Despite the omnipresence of discussions of cleft and pseudo-cleft sentences in English in much of the linguistic literature of the past two decades, Collins' investigation, which is based on the London-Lund (LL) and Lancaster-Oslo/Bergen (LOB) corpora, is the first major corpus-based study of these constructions.

Collins deals with most aspects – syntactic, semantic and pragmatic – relevant to clefts and pseudo-clefts, though his main concern is with their discourse functions and informational/thematic characteristics. His basic contention is that each of the construction types examined has a distinct combination of logico-semantic, thematic and informational properties and thus a unique communicative value. On the basis of this broad corpus investigation, Collins is able to offer ample empirical support for some of the claims concerning these constructions presented in previous studies and to refute others, as well as to contribute with a number of new perspectives. The approach adopted is functional, largely within the framework of Halliday's systemic-functional grammar. Collins draws on Halliday's and Huddleston's textually based analyses of the constructions in question, and he also acknowledges his debt to Prince, whose seminal article (1978) was of major importance in drawing attention to the informational properties of clefts and pseudo-clefts.

I find it convenient to begin with a summary of the book (Sections 1–3) and to deal separately (Section 4) with some of the points of criticism to which it gives rise.

1. Preliminaries

1.1. The objects of investigation and the corpora

The book is divided into eight chapters, followed by a combined author and subject index. The brief introductory chapter offers a first presentation of the types of construction dealt with as well as a brief outline of a functional approach to the study of these constructions.

The terminology proposed by Collins may be summarized as follows: sentences like (1) and (2) are called *clefts* and *pseudo-clefts* respectively:

(1) It was a sherry that Tom offered Sue. (p. 1)[1]

(2) What Tom offered Sue was a sherry. (p. 1)

The second clause of clefts and the initial clause of pseudo-clefts like (2) are referred to as *the relative clause*,[2] and the postcopular constituent as *the highlighted element/item.* Clefts and pseudo-clefts are *identifying constructions*, expressing a relationship of identity between the relative clause as *identified element* and the highlighted element as *identifier.* Pseudo-clefts like (2), where the relative clause/identified element is subject, are labelled *basic pseudo-clefts*, whereas pseudo-clefts with the highlighted item/identifier as subject are called *reversed.*

The relevant aspects of the two corpora used as bases for the investigation are presented in Chapter 2. The fact that LOB and LL are machine-readable must have been less of an asset in this study than in many others, since clefts are notoriously hard to identify on the basis of syntactic and morphological criteria alone. In fact, Collins has identified the cleft sentences by highlighting and then manually screening all occurrences of *it* and all forms of the verb *be*.

1.2. Definitions and delimitations

Chapter 3 is devoted to defining and delimiting the classes of clefts and pseudo-clefts. In addition to a thorough discussion of attributive constructions that are superficially similar to, but in reality distinct from, the constructions under investigation, Collins deals with a number of borderline cases.

He adopts a wide definition of **pseudo-clefts**: in addition to the prototypical (and statistically dominant) ***wh*-clefts**, as in (3), this class

includes ***th*-clefts** and ***all*-clefts**. ***Th*-clefts** are identifying constructions with relative clauses introduced by expressions like *the thing (that)*, *the one (who)*, *the place (that/where)*, etc., as in (4). ***All*-clefts** are included on the basis of the semantic similarity of sentences like (4) and (5):

(3) What they took was her purse.

(4) The only thing they took was her purse.

(5) All they took was her purse.

The discussion of *th*-clefts centres on two problems of delimitation: what nouns may be regarded as 'pro-form equivalents' of the interrogatives, and the extent to which the lexical head may be modified. Drawing upon a Hallidayan distinction between general nouns and pro-nouns, Collins includes the pro-nouns *one*, *place*, *thing*, *way* but excludes general nouns like *man*, *person*. Thus sentences like (6) are included in this study, whereas those exemplified in (7) and (8) are not:

(6) Frank Morgan was the one who started all this.

(7) Frank Morgan was the person who started all this.

(8) Frank Morgan was the man who started all this.

With regard to modification, Collins would accept sentences like (9) and (10) as pseudo-clefts, but not (11), the test being whether a non-cleft version is available which has the same propositional content (cf. *They did that first*, *They only did that* vs. **They did that best*):

(9) That was the first thing they did.

(10) That was the only thing they did.

(11) That was the best thing they did.

In the class of **clefts**, Collins includes so-called predicational clefts like (12), and sentences with expressions like *it may be that*, *maybe it is that*, *can it be that*, *it isn't that*, as in (13), in addition to the prototypical type illustrated in (1) above. Like Declerck (1988), Collins argues convincingly for the exclusion of proverbial sentences like (14), which a number of scholars have regarded as clefts:

(12) It is not a sentimental, but a precise point which he makes: (p. 41; LOB G59, 93-4)

(13) it may be that a frontal view will be more effective in certain circumstances (p. 35; LOB E10,92-3)

(14) It is a poor heart that never rejoices. (p. 40)

A more extensive delimitation problem is represented by copular expressions introduced by *it* + *be*. Collins counts as elliptical clefts constructions where the elided relative clause may be recovered 'either directly from the co-text or context, or indirectly via inferences from them' (p. 46), but excludes, among others, copular sentences with a human referent as predicate nominal, *it's John*, etc. These, as well as other problems of delimitation, will be discussed in further detail below (Section 4.1).

1.3. Syntax and semantics

In the fourth chapter, entitled 'Formal properties', the author's main concern is with the function of the highlighted item. The chapter also contains a critical presentation of various syntactic analyses and a section dealing with semantic properties (identification, exclusiveness implicature, presupposition). Both are rather brief, considering the vast attention that has been devoted to these phenomena in the linguistic literature, but most of the salient points of previous discussions are dealt with. Moreover, the focus of attention in this book is professedly on other aspects of the constructions in question.

2. The communicative approach

In Chapter 5, 'Communicative meanings', Collins develops his framework for the textual analysis of clefts and pseudo-clefts in terms of their thematic and informational characteristics. The former refer to the organization of the sentence into a thematic and a rhematic part and the latter to the distribution of 'new' and 'given' information. Chapter 6 describes the results of the application of this theoretical framework to the corpus data. The findings are carefully illustrated with examples, some of which are discussed in great detail.

2.1. Theme

Theme is defined in Hallidayan terms as 'the point of departure of the clause', which in English is realized by the initial item(s), i.e. the highlighted item in the case of clefts and reversed pseudo-clefts and the relative clause of basic pseudo-clefts. Clefts and pseudo-clefts, Collins points out, divide the sentence into two distinct parts which explicitly distinguish between the theme and the rheme. The pseudo-cleft constructions, it is suggested, are 'thematic resources'; their *raison d'être* is to enable virtually any element to be thematized.

One of the basic claims of this study is that the thematic prominence or highlighting of (reversed) pseudo-clefts differs semantically from that of clefts. In reversed pseudo-clefts, where the theme is rendered prominent by 'representing it as one of the two members participating in an equative relationship' the prominence is said to be 'ideational' or 'experiential', whereas in clefts, where the theme is structurally predicated and the theme-rheme structure is not reversible, it is 'essentially textual' (p. 84).

2.2. Information

Information is chiefly defined with reference to prosodic features. The discourse is organized into tone units, in which one constituent, marked by the intonation nucleus, is focal, representing new information. Givenness, on the other hand, is signalled prosodically by absence of prominence. However, a givenness concept that is primarily related to prosody is obviously problematic in the description of written texts, and the information structures of clefts and pseudo-clefts are therefore also described in terms of recoverability. Thus given information is defined, with Halliday, as what the speaker presents as recoverable from the co-text or context. Such information, it is argued, will normally be either previously mentioned, generally known or physically present. New information, on the other hand, is defined as information that is either introduced for the first time or treated differently when mentioned a second time, e.g. in the case of contrastive emphasis. (See Section 4.2 below for further discussion.)

2.3. Basic pseudo-clefts

In basic pseudo-clefts, the relative clause serves grammatically as the subject, semantically as the identified element. Moreover, Collins suggests, it is the theme at the textual level, it expresses a presupposition at the logico-semantic level and is primarily associated with given information. The rhematic highlighted item, grammatically the subject complement, conveys the new information.

The data from LL indicate that basic pseudo-clefts normally consist of more than one tone unit, i.e. the thematic relative clause is realized as a separate tone unit, as in (15):

(15) # – it's just that – – – what they KN\ /OW about # is experimental RES\EARCH # (p. 118; LL S.2.4,727-8)

According to the prosodic definition of information, the relative clause

consequently contains some material marked as new, viz. the focus of information indicated by the nucleus of the tone unit. However, this newness, it is claimed, is attenuated owing to the fact that the clause is also thematic, presupposed and syntactically dependent (i.e. it is a subclause), and therefore has a 'backgrounded character'.

Contra Prince (1978) and Gundel (1985), Collins argues that it is the specific combination of these discourse variables, and not a specific type of givenness, that accounts for the 'special communicative flavour of the relative clause of basic pseudo-clefts, in which the speaker appears to be making assumptions about notions that are, or could be, in the hearer's consciousness' (p. 100). The conflation of theme, presupposition, givenness, syntactic dependency and 'identified' status serves, he suggests, to persuade the addressee that s/he should be able to recover, and hence accept, the material presented in the relative clause. The strong sense of givenness also explains why these constructions rarely occur discourse-initially.

Basic pseudo-clefts are thus claimed to be 'givenness-oriented', and Collins proposes a detailed taxonomy for the kind of given information that appears in these constructions (pp. 95-97): four co-textual types, derived from the parameters direct/indirect recoverability and similarity/oppositeness, and three contextual types (the chief exponents of which may be exemplified by pseudo-clefts introduced by *What happens is...*, *What worries me is...*, *What I mean is...*).[3]

Not unexpectedly, basic pseudo-clefts with contextual antecedents are found to be more common in speech, whereas co-textual antecedents are more common in writing.

2.4. Reversed pseudo-clefts

In reversed pseudo-clefts the highlighted element is theme and the relative clause is rheme. They are normally realized as a single tone unit, with the nucleus in the rheme/relative clause as in (16):

(16) John's wife has left him and # *that's why he's UPSET* # (p. 139)

As in the case of basic pseudo-clefts, the relative clause thus contains elements prosodically marked as new information. But here too, the fact that it is syntactically subordinated and represents a presupposition is claimed to weaken the newness, so that 'the information is presented as something which is not-at-issue, something on which doubt or disagreement is not countenanced' (p. 97).

The theme/highlighted item of these constructions is normally a demonstrative pronoun – often anaphoric and therefore inherently given.

Collins makes the interesting observation that this construction contains little new information, but serves typically to relate a 'deictically-referred-to stretch of previous discourse' (p. 146) with something that is presented as background material. This summing-up function, it is observed, renders the reversed pseudo-cleft particularly suited to marking the conclusion of stages in the discourse.

2.5. Clefts

Clefts are classified informationally in terms of combinations of four categories of informativity, somewhat heterogeneously defined in terms of prosodic salience, syntactic features (subordination, pronominalization, deixis, etc.) and, primarily, textual evidence (recoverability). The highlighted item and the relative clause enter into different combinations of information that is either fresh/new[4] (not recoverable), contrastive (recoverable, but freshly attended to), stale/given (directly recoverable), or inferable (recoverable by inference).

Clefts where the highlighted item represents either new or contrastive information, while the relative clause represents information that is given or inferable, are called **Type 1 clefts**. They are also referred to as **unmarked** clefts. This is the basic, well-known type, corresponding to Prince's (1978) *stressed focus* it-*clefts*. Collins' findings indicate that Type 1 clefts are normally 'realized as a single tone unit with focal highlighted element' (p. 159), i.e. the relative clause is prosodically marked as 'given'. The low informativity of the relative clause provides an explanation for the fact that it may be elided, as in (17):

(17) It is not the observation of likenesses which is at fault in popular etymology, *it is the fact that conclusions about the relationships of words, drawn from comparisons, happen to be erroneous.* (pp. 46, 160; LOB G51,59-61)

Clefts where it is the second clause that conveys the 'news', *informative presupposition* it-*clefts* in Prince's terminology, are considered **marked**. Collins' corpus data reveal that this group of clefts is not as homogeneous as suggested by Prince. He distinguishes between two types of marked clefts: in **Type 2** clefts the highlighted item is given or inferable, whereas the relative clause represents the new or contrastive information. Typically, the highlighted item is short and anaphoric or deictic:

(18) It should be remembered that until the implementation of the Guillebaud Report, under which railway rates of pay were based on the principle of 'comparability' with those of comparable em-

ployees in other employments, railwaymen had worked for considerably debased rates of pay, and *it was they who had been providing the subsidy necessary for the running of the railways which are necessary to the economy of the country.* (Collins 1991:511; LOB B11,56-8)

In **Type 3** clefts both the highlighted element and the relative clause convey new or contrastive information. However, the 'bulk of the propositional content' is found in the relative clause, while the highlighted element tends to be a circumstantial or scene-setting adjunct of time, place or the like, as in (19):

(19) It was not long ago that Richard Rodney Bennett composed a 'Calendar' for chamber ensemble. (p. 167; LOB C01,73-4)

This type, unlike Types 1 and 2, is frequently used discourse-initially. This is convincingly argued to be a reflection of the fact that on the one hand, Type 3 clefts primarily convey new information, yet on the other, the cleft construction gives the information 'a character of non-controversiality', serving to 'give the impression that the listener/reader is simply being "put in the picture", or "brought up to date" with information to which others will already be privy' (p. 166). Thus the cleft construction serves to 'moderate the brusqueness which might result, in the corresponding non-cleft, from the presentation of unmitigated new content in topic-sentence position' (p. 166). This is reminiscent of arguments put forward by Prince (1978), but as in the case of pseudo-clefts, Collins argues that these characteristics are derived not from a specific type of givenness, but from a particular configuration of syntactic, semantic, thematic and informational properties.

3. Mode and register variation

Chapter 7 explores the distribution of clefts and pseudo-clefts in speech and writing and in the various registers. The findings are carefully tabulated and described and explained with reference to the characteristics established in previous chapters. Only a few of the numerous observations presented in this chapter will be mentioned here.

3.1. Clefts in speech and writing

Clefts outnumber pseudo-clefts in LOB, whereas the reverse obtains in LL. The relative popularity of clefts in writing must primarily be attributed to the marked varieties, since Type 1, i.e. unmarked clefts, are reported to occur more frequently in the spoken corpus. They are,

moreover, favoured in informal registers. Type 3 clefts, on the other hand, are 'preferred in formal, learned writing', a preference that is explained in terms of their 'high level of informativity'.

3.2. Pseudo-clefts

Pseudo-clefts were found to be far more frequent in the spoken than in the written corpus. This distribution, Collins argues, can be explained with reference to their textual characteristics. **Basic pseudo-clefts** allow the speaker to explicitly specify background knowledge before presenting the 'message'; thus they function as indices 'within the flow of speech' (p. 181). **Reversed pseudo-clefts** are used more frequently than their basic counterparts in both speech and writing, which is explained in terms of their summing-up or 'internal referencing' function.

4. Some critical comments

4.1. Problems of delimitation

The question of delimitation and identification may seem to be an essentialist pseudo-problem (in the sense of Janicki 1989) to anyone content to describe the prototypical aspects of the canonical cleft or pseudo-cleft, but a discussion of this question has its obvious place in a study that relies extensively on quantitative analyses. Although one may not agree with all his decisions, Collins is to be applauded for taking this problem seriously, as well as for his willingness to call attention to 'troublesome data'.

As was noted in Section 1.2 above, one such problem of delimitation is encountered in the case of copular expressions introduced by *it* + *be*, which Collins has chosen to regard as elliptical cleft constructions if the elided relative clause may be recovered 'either directly from the co-text or context, or indirectly via inferences from them' (p. 46). Cases like (17) above, where the relative clause may be directly recovered from the context, are uncontroversially cleft. However, it is the kind and degree of inferential recoverability that is problematic. With copular sentences of the kind referred to here it is very often possible to think of a relative clause extension that is more or less plausible in the co-text or context. Still, Collins has chosen to be restrictive rather than admit doubtful cases. He therefore excludes copular sentences with a human referent as predicate nominal, *it is me*, *It's John*, etc. Arguably, however, such sentences are more amenable to a cleft interpretation in some contexts than in others. Thus, *It's John* in answer to *Who's singing?*

may be more plausibly expanded to a cleft (*It's John who's singing*), than in answer to *Who's that?* (**It's John who that is*).

Following Declerck (1981; cf. also Declerck 1988: 144-45), Collins also excludes sentences where *it* refers to a sense perception of e.g. a noise, or to a general notion like trouble, thing, reason, cause or question. The italicized portion of example (20) below is therefore rejected as an elliptical cleft on the grounds that the complete form would be *it was the cat mewing*, where *it* refers to an implied nominal – *the noise*:

(20) 'The cat will have got itself out through the coal-shoot. Bound to –'

'It hasn't. I heard it mewing. I am sure and certain *it was the cat* – let go of me, George!' (p. 47; LOB P01,105-6)

It is not difficult to accept that the interpretation suggested by Collins, like the examples discussed by Declerck, is unambiguously non-cleft. However, it is doubtful whether reference to sense perception is *per se* incompatible with a cleft interpretation. In the case of (20), the following (cleft) interpretation is just as conceivable: *it was the cat that I heard.* On this point, then, Collins appears to have been overly restrictive. Sentences like (21), on the other hand, he has interpreted as two clefts, rather than in terms of a right-dislocation of part of the highlighted item. This is somewhat surprising in view of the fact that contrastiveness is part of the communicative meaning of many clefts, and contrasts may be explicitly introduced as part of the highlighted item, as in (22):

(21) It's their interest you want – not their sympathy. (p. 46; LOB F03,175-6)

(22) It's their interest, not their sympathy, you want.

This is, then, an area where there is considerable room for individual interpretation, and it is therefore extremely difficult to draw a hard and fast line between elliptical clefts and certain other copular constructions. This problem (which is, of course, implicitly acknowledged in Collins' discussion), together with the fact that copular sentences of the *it's me* type are highly frequent, particularly in speech, provide somewhat shaky grounds for statistically based conclusions.

A similar objection concerns the inclusion in the group of clefts of constructions with expressions like *it may be that, maybe it is that, can it be that, it isn't that*, as in (23) and (24) below:

(23) # [ə] it's not that Mervyn's [ə$^{?}$ə$^{?}$] T\ /OTALLY unreliable # (p.

34; LL S.2.6,119)

(24) # [i] is it that you you're not in a different P\OST TH/ERE # (p. 35; LL S.2.7,410)

Sentences of this type are variously referred to by Collins as clefts without 'a highlighted element with experiential function' (p. 34) – also called 'ideational function' and 'representational function' – and as clefts with 'zero-highlighted item' and clefts with 'zero theme', presumably depending on the perspective. Judging from the examples quoted, such sentences are probably also what is meant by the expression clefts 'without topical theme' (p. 158).

'Within this category', it is argued,

> the relative clause contains all items having an ideational function in the sentence, the cleaving serving to highlight non-ideational items relating to tense, modality, aspect and polarity. (p. 57)

In other words, the focus is on the temporal, modal or aspectual categories carried by *be* and modal auxiliaries, and, presumably, on the contrast implied by the negative adverb, as in (23) above.

Collins admits that the claim that sentences such as these 'are in fact clefts rather than structures deriving from extraposition of the nominal clause requires some justification', and furthermore, that 'convincing evidence is difficult to find' (p. 35). This is hardly surprising to anyone who is convinced that they are not appropriately classified as clefts.

As purported evidence in favour of the cleft analysis, Collins points to the 'typical availability of a non-cleft counterpart for such sentences' (p. 35), as in (25) and (26):

(25) Mervyn's not totally unreliable. (p. 35)

(26) Are you not in a different post there? (p. 35)

He concedes that uncleaving is not possible with all the relevant sentences, viz. in sentences with multiple auxiliaries (which, he suggests, are problematic in uncontroversial clefts too). However, it may be added that this uncleaving test does not usually work with adverbs other than *not* either. The sentence in (27) cannot be paraphrased as (28), and (30) does not preserve the propositional content of (29):

(27) It's only that Mervyn's totally unreliable.

(28) *Mervyn's only totally unreliable.

(29) It's partly that Mervyn's unreliable.

(30) Mervyn's partly unreliable.

Collins also refers to the lack of a non-extraposed counterpart, as in (31) and (32), thereby suggesting that (23) and (24) are not extraposition structures:

(31) *That Mervyn's totally unreliable is not. (p. 35)

(32) *Is that you're not in a different post there? (p. 35)

However, it is hardly necessary to point out that verbs like *appear, seem, happen, turn out* do not admit subject that-clauses either, a fact that is usually accounted for in terms of obligatory extraposition. Compare the extraposition structure in (33) with the ungrammatical non-extraposed variant in (34):

(33) It seems that Mervyn's unreliable.

(34) *That Mervyn's unreliable seems.

In fact, sentences like (23) and (24) are mentioned by Quirk et al. (1985: 1392 n) among the constructions with obligatory extraposition of the *that*-clause. This is not the place to undertake a detailed analysis of the constructions in question, but it seems that they are indeed best analysed in terms of extraposition,[5] and that *be* in these sentences is not the copula, but the lexical verb meaning 'to be the case or the fact' (cf. OED, sv *be* B.I.2-3).

Finally, Collins refers to what he calls 'thematic evidence in favour of a cleft interpretation' (p. 36), which he acknowledges to be 'suggestive, rather than decisive'. The 'non-ideational item(s) following the *it*' are claimed to be thematically prominent, which, it is argued, suggests a cleft interpretation, since the primary function of clefts is to assign thematic prominence, whereas the communicative function of 'extraposition structures is generally interpreted in terms of the principle of "end-weight"' (p. 36). It is difficult to see, however, that there is a difference in thematic prominence between a purported cleft candidate like (35), and a sentence like (36), which, as noted above, is commonly described in terms of extraposition.

(35) It may be that Mervyn's unreliable.

(36) It may seem that Mervyn's unreliable.

Obviously, the last word has not been said about these constructions, and Collins is right in suggesting that they are rather elusive. However, the conclusion must be that their syntactic, semantic and pragmatic features do not warrant an inclusion in the group of clefts. Thus Delahunty's (1984: 65) proposition, challenged by Collins, that 'The clause [the relative clause in Collins' terminology] always contains a

gap, or trace, with which the focus [highlighted item] is (ultimately) associated' is still valid.

The justification for discussing the purported zero-type clefts at such length is that they account for 14.5% (109 tokens) of the total number of clefts (752) in the two corpora, and that zero is listed as the third most common type of highlighted item in clefts (cf. pp. 55-57). The inclusion of this type of construction therefore has obvious consequences for some of the conclusions in a study that relies heavily on statistical analyses of frequency. In some cases it is possible to single out the alleged 'zero-highlight clefts' and simply deduct them from the reported figures. In other cases this is not feasible, and it is difficult to decide whether their exclusion would yield significantly different results. Besides the statistical aspect, it should also be noted that the very existence of such clefts is used to underpin one of the central claims of this book, namely that clefts and pseudo-clefts display different types of thematic prominence – textual/predicational and ideational/equational respectively. Furthermore, it is argued that one of the properties that account for the popularity of clefts in writing is 'the "paradigmatic" thematic flexibility of the cleft construction, which readily permits prepositional phrases and zero as the highlighted element' (p. 215). This observation anticipates the discussion in the ensuing section, in which some further explanations for the distribution of clefts according to mode (cf. Section 3.1) will be briefly attended to.

4.2. Miscellaneous observations

Collins indicates a number of different reasons for the relative popularity of clefts in writing. One very plausible explanation relates this fact to the dense 'information-packing' and high degree of communicative dynamism of marked clefts. Less convincing, however, is the suggestion that this distribution may also be attributed to the newness associated with the theme/highlighted item of both Type 1 and Type 3 clefts. This proposal is incongruous with the observed prevalence of Type 1 clefts in speech, and with the fact that the newsworthiness of the highlighted item of Type 3 clefts is after all of subsidiary importance, described as 'perhaps more appropriately "semi-new"', than new (p. 205). In the same vein Collins suggests that it is

> perhaps as much on account of their thematic flexibility, as it is their unambiguous indication of the intonationally focal constituent, that clefts are so popular – relative to pseudo-clefts – in written discourse (p. 174).

The thematic flexibility refers to the claim, mentioned at the closing of the preceding section, that the cleft construction 'readily permits prepositional phrases and zero as the highlighted element' (p. 215). However, the 'zero' constructions are in fact reported to be relatively more numerous in LL (cf. Table 7.9). But more importantly, as I argued in Section 4.1 above, they are not easily analysed as clefts.

The idea that clefts are a syntactic means of assigning intonational prominence to the highlighted item is a commonplace in the text-book literature. However, Collins' data reveal that in speech the unmarked pattern, where the highlighted element is focal, is in fact 'commonly overridden' (p. 155). The 'unambiguous indication' of the focus attributed to the cleft construction therefore depends crucially on the 'latent unmarkedness of intonation in writing' (p. 189), i.e. the assumption that written texts will normally be assigned unmarked intonation.

Yet another characteristic which according to Collins contributes to the comparative popularity of clefts in writing is their structural similarity to constructions of the type *it is said that..., it is well-known that....* From these, it is claimed, clefts 'derive a depersonalized quality and a formality' that is incompatible with informal speech (p. 215). It is true that Type 3 clefts are commonly used in texts where there is also a high frequency of impersonal passives and extraposition structures. However, the primary explanation for this distribution seems to be that the type of text that favours structures which permit agent suppression also tends to have constructions with precisely the sort of 'dense information-packing' that marked clefts allow, and which is indeed uncharacteristic of informal speech.

An entirely different objection concerns the value of a prosodically based analysis of the information structure of clefts and pseudo-clefts. In Section 2.2 above it was pointed out that the concept of information is defined partly in terms of prosody, partly with reference to the textual notion of recoverability. The prosodic approach has supremacy in the theoretical framework as presented in Chapter 5: 'In this study I shall adopt a primarily Hallidayan approach and regard nucleus placement within the tone unit as the main determinant of information structure' (p. 90). The concept of recoverability is however prevalent in the discussion of corpus data. In the presentation of examples from LL, a conflict is repeatedly observed between the two approaches. Information that is 'given' from the point of view of recoverability is not infrequently presented in a separate tone unit, which by definition contains a nucleus. And according to the prosodic definition of information, the nucleus is associated with new information. As was pointed out in Section 2.3

above, this is in fact the normal state of affairs in basic pseudo-clefts. Their thematic part is a clause, and Collins' data indicate that it is almost invariably realized as a separate tone unit, and therefore has an element containing new information. On the other hand, the author consistently argues that basic pseudo-clefts are strongly givenness-oriented. He attempts to account for this contradiction by suggesting that the newness indicated by the nucleus is 'attenuated' by the thematic, syntactic and semantic features of this construction. It is even contended that this opposition is part of its particular communicative value:

> In fact it is the tension that results from the conflation of apparently incongruous elements representing different linguistic systems which, in conjunction with other mappings, gives rise to the unique communicative meaning generated by the basic pseudo-cleft construction (pp. 120-21).

Collins also observes that the nucleus normally falls on an item that is retrievable, and that the prosodically indicated newness therefore is 'contrastive rather than fresh', and that it is '"new" in a mildly contrastive sense'. It is further claimed that the 'meaning of the information focus is here simply "pay attention to this item, because I believe that it requires special emphasis"' (p. 120).

This auxiliary hypothesis, introduced to save 'the theoretically significant generalization that nucleus-bearing items represent new information' (p. 220 n), is not entirely convincing. It is even less convincing in the light of examples like the following, where the contextually contrastive part is not prosodically marked as contrastive, while the focus is on an item that is not informationally salient from a textual point of view:

(37) #. I P/ERSONALLY # would not trust . the German N=ATION # . as I've known it during my L/IFETIME # . with a P\OPGUN # let alone with [ði:] R/ \EAL machinery # of W\AR # – *what I would trust them W/ITH # . is the machinery of P\EACE* # (p. 124; LL S.5.1,689-90)

It is hard to imagine what sort of 'newness' is to be associated with the (contextually given) focal item, *W/ITH*. It is not even 'mildly contrastive', nor is it an item that the speaker would conceivably want to indicate as especially important. The italicized part of (38) is a parallel cleft example:

(38) # . Scottish could you see be S\EEN # as a service \INDUSTRY ((in)) { some . RESP\ECTS # } # . [ə:m] *it's not that aspect I'm \AFTER* # it's the fact that there's a fair N\UMBER of us # in

this SC\OTTISH department # . WH\O #. might just as W/ \ELL # be doing the things we're very G\ /OOD at # in other sorts of DEP\ARTMENTS # (p. 156; LL S.3.4,863)

Examples like (37) and (38) are atypical, yet they suggest that in cases of conflict between a prosodically based analysis of the information structure in clefts and pseudo-clefts and an interpretation in terms of recoverability, it is the latter that prevails.

Collins rejects Taglicht's (1984: 42) suggestion that 'All "new information" is represented by focal items, but not all focal items represent "new information"' (p. 220). However, this does indeed seem to be an adequate description of the situation in most basic pseudo-clefts as well as in examples like (37) and (38). It would appear, then, that the information structure of clefts and pseudo-clefts is best described in terms of recoverability, while a concept of information defined in terms of prosody does not contribute as significantly to the understanding of the discourse properties of these constructions.

My final comment concerns the concept of markedness, a recurrent feature in systemic grammar, but here used to excess. The terms *marked* and *unmarked* are applied to such a variety of phenomena – informational, structural, thematic and intonational – and used at so many different levels that the reader is sometimes bewildered. The discussion of the intonational features of clefts on pp. 154-58 is particularly impervious in this respect, but a less complicated example will do as an illustration: Themes are normally considered (informationally) unmarked when associated with given information, so also in the case of basic pseudo-clefts. In clefts and reversed pseudo-clefts, however, the combination theme – new information is considered unmarked. Thus when it is said, for instance, that clefts display 'unambiguous mapping of theme on to new information in the unmarked instance (for which there is preference in writing)' (p. 175), the reader should keep in mind that this does not refer to (structurally) unmarked clefts (which, as noted above, are more common in speech than in writing), but to a combination of unmarked and marked (Type 3) clefts, both of which have themes that convey non-recoverable information. It should also be remembered that *unmarked* normally refers to the most usual realization of a particular variable, but that this is not always the case. For specific reasons (that are carefully explained), reversed pseudo-clefts with a focal highlighted item are regarded as unmarked, despite the fact that in 95 % of the cases the nucleus is reported to be, not on the highlighted item, but in the rheme/relative clause/identified element.

5. *Concluding remarks*

A book so rich in detail as the present one is an important source of information and a stimulus to further research. But at the same time it lays itself open to criticism: the presentation is at times repetitious and inconsistencies occur. All in all, however, the shortcomings that have been pointed out do not seriously detract from the value of this impressive study, which is full of insights and keen observations; in fact, it is one of the most stimulating accounts of cleft and pseudo-cleft constructions that have been published to date. Collins' approach is indeed 'fresh', and his book is a welcome addition and corrective to the existing literature.

Notes

1 The reviewer has been slightly more liberal than the author in introducing what the latter refers to as 'invented' or 'contrived examples'. Thus examples that are not provided with a page reference do not appear in the book. It should also be noted that for reasons of space, the contexts of examples from LOB and LL have generally been curtailed or omitted. The transcription symbols are those used by Collins, with the following exceptions: rising or falling nuclus is indicated by a slash (/) or backslash (\), respectively, preceding the accented vowel, and (rising–) falling–rising or (falling–) rising–falling nucleus by the combinations (\ /) and (/ \). An equals sign (=) is used for level nuclear tone.

2 This terminology is also used in the case of clefts with an adjunct or a clause as highlighted element, reflecting a desire to account for the second clause in clefts 'within a single analysis' (p. 53). It is viewed, that is, as a specific brand of relative clause that, among other deviations from the characteristics of ordinary relative clauses, may have an adjunct as 'antecedent'.

3 An unfortunate editing slip must have occurred in Section 5.3.4 (pp. 95-97), where Collins' model for givenness in basic pseudo-clefts is presented. As indicated above the distinction between two types of givenness, co-textual and contextual, is central to this model. The former type is discussed and exemplified at length, but there is no mention of the latter beyond its inclusion in Table 5.1 (p. 96). In a recent article in *Linguistics* (Collins 1991, which, incidentally, is an excellent summary of some of the main points of the book), there is however a page-long passage where contextual givenness is

amply described and exemplified. The passage (pp. 502-503) is found in a section of the article that is otherwise nearly identical to Section 5.3.4 of the book. That the omission of this passage is indeed inadvertent is also indicated by the fact that reference is made to such a discussion in Chapter 6, where this model is applied to corpus examples.

4 The terms *fresh* and *stale* are rather confusingly used both as synonyms and hyponyms of *new* and *given* respectively.

5 Or at any rate, as parallels to constructions that are commonly described in terms of extraposition. It may be noted that historically (and from certain other points of view) the 'extraposed' variants are more 'basic' than their non-extraposed counterparts: nominal *that*-clauses which in PresE appear in so-called non-extraposed subject position were not used in this position in OE. There is, for instance, no OE parallel to sentences like *That he's unreliable is true* (cf. Visser 1963: 19, Mitchell 1985: 11).

References

Collins, Peter Craig. 1991. Pseudocleft and cleft constructions: a thematic and informational interpretation. *Linguistics* 29:481-519.

Declerck, Renaat. 1981. Pseudo-modifiers. *Lingua* 54:135-63.

Declerck, Renaat. 1988. *Studies on copular sentences, clefts and pseudo-clefts*. Leuven: Leuven University Press.

Delahunty, Gerald. 1984. The analysis of English cleft sentences. *Linguistic Analysis* 13: 63-113.

Gundel, Jeanette K. 1985. 'Shared knowledge' and topicality. *Journal of Pragmatics* 9:83-107.

Janicki, Karol. 1989. A rebuttal of essentialist sociolinguistics. *International Journal of the Sociology of Language* 78:93-105.

Mitchell, Bruce. 1985. *Old English syntax*. Vol II. Oxford: Clarendon.

Prince, Ellen. 1978. A comparison of WH-clefts and *it*-clefts in discourse. *Language* 54:883-906.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1985. *A comprehensive grammar of the English language*. London: Longman.

Taglicht, Josef. 1984. *Message and emphasis: On focus and scope in English*. London: Longman.

Visser, F. Th. 1963. *An historical syntax of the English language*. Vol I. Leiden: E.J. Brill.

The Nijmegen Linguistic Database program

Clive Souter
University of Leeds

Hans van Halteren and Theo van den Heuvel. *Linguistic exploitation of syntactic databases: The use of the Nijmegen Linguistic Database program.* Amsterdam: Rodopi, 1990. 207 pp. ISBN 90–6203–809–3.

The last five years have thankfully seen a significant increase in the development and use of English corpora for computational linguistic work, as researchers seek to test the performance of their grammatical models against large, putatively representative, collections of text. This rise in use and interest was envisaged by van Halteren and van den Heuvel in the introduction to this book, with particular respect to parsed (fully grammatically annotated) corpora. The book describes a software tool for the storage and exploration of parsed corpora, or treebanks, as they have become known elsewhere. The tool goes by the rather unremarkable name of the Linguistic Database (LDB). In fact the book is accompanied by a demonstration copy of the LDB software on diskette, so this review will cover both the book and the software.

1. The LDB book

The book, or the user manual as it might be called, is divided into three parts; a very brief introduction, a tutorial section, and a reference section. The division into a tutorial and a reference section is a neat way to try and kill two birds with one stone. The tutorial contains a few suggested exercises, for which some solutions are to be found in an appendix. A further appendix outlines machine-dependent commands and keystrokes for loading and running the software, and is followed by a short bibliography and a very useful index. The index is so important because the reference section is not ordered alphabetically by

command name, but instead takes you through the software in more detail. The text is generously illustrated with example menus, tree views, and exploration schemes, which are valuable in helping the reader's understanding of the text, and therefore the software.

The fact that neither of the authors[1] is a native English speaker goes largely unnoticed, although the text does include occasional spelling and grammar errors, (eg. p.4: reversely, p.5: form, p.5: information ... are captured, p.21-22: repeated line, p.26: wil, p.92: Appendix). Section 4.12, Advanced Use of the LDB, appears from the text to have been intended as a separate chapter, which would indeed provide a better structure. These inaccuracies detract from what is otherwise a well presented text, in which the authors generally explain their subject clearly. The need for such a book goes without saying, given the existence and distribution of the software. The content of the book then, is a detailed explanation of the LDB software, which forms the main part of this review.

2. The LDB software

The software demonstrator contains a small part of the Nijmegen Corpus: 200 sentences taken from the fully annotated 130,000 word corpus of mainly written English (Keulen 1986). Users of the demo are able to browse through this subset of the corpus and conduct searches for patterns of particular interest, rather as they would with the whole corpus. The authors have written the software specifically for the storage and search of syntactically analysed corpora, because existing database systems either could not handle large numbers of tree structures, or if they could, did so in an inefficient manner. Other tools exist for the handling of computer corpora (Oxford Concordance Program, WordCruncher, TACT) but only for raw or word-tagged text. The LDB is, to my knowledge, unique in the fact that it handles fully parsed corpora. It is available in versions to run under MS-DOS, Unix and VMS Operating Systems on IBM-compatible PCs, Sun and SPARC workstations, and VAX mainframes. The full copy of the software (instead of the demonstration) is available from the authors. The demonstration version is loaded straightforwardly, and can be run in three modes, short demo, detailed demo and interactive demo. The first two walk you through the system functions, and the interactive mode allows you to take control.

To evaluate the usefulness of the software, we might ask what sort of functions users would want from a database tool for parsed corpora. What the LDB offers:

2.1 Browsing

The usual facility of browsing through the corpus is supported by a tree viewing system with two choices: *Map view* shows the basic tree structure with nodes simply numbered to identify common sisters, but normally displays most or all of the tree. *Environment view* shows a localised subtree complete with the grammatical labels (syntactic class, function, and features, if any) of each node. It is one of the disadvantages of the system that the size of most trees forces this binocular view of what would normally be perceived as a single structure. The selection of a subpart of the whole tree for particular attention is achieved using the focus, which the user can move about from node to node with a variety of keystrokes. It is a shame that moving the focus about the tree needs to be so long-winded, when most machines now support a mouse, which would simply allow the user to click on the node to be selected. This feature is not included because the software was actually written several years ago. It is also possible to scroll up and down, and to the left and right, in order to view a large tree. The keys U,D,L and R serve this purpose, but in what appears to me to be a counter-intuitive way: the U key takes you down the screen and the D key takes you up; the L key takes you right, and the R key takes you left! Presumably the perception is that it is the tree that is being moved around, rather than the reader's view of the tree, but this set-up goes against all practice I have seen in screen editors and word-processing packages elsewhere.

2.2 Searching

The LDB provides exploration schemes to allow users to search for patterns of particular labels and structures from the whole corpus. Such schemes consist of two parts. First, a pattern to search for, and second, an activity to perform with all the trees that match the pattern. A fairly rich logic is provided to allow combinations of patterns. The activity part of these schemes might have been better dealt with by sending the output of the pattern to the screen by default, with an option for writing to a file. If the user neglects to choose an activity to be performed with every successful match, it appears as if no matches have been found at all.

2.3 Dividing up the corpus

Users can easily subdivide the corpus for any purpose, such as comparison of the structures pertaining to different text genres, or indeed to store the output of an exploration scheme.

2.4 Input and output

ASCII files can be created with output from various processes in the LDB, including the tree viewer, for subsequent printing. Input can be conducted in character or line mode, for reading in items such as filenames.

2.5 Selecting and deleting a corpus

If more than one corpus is available in the database, the LDB requires you to choose which is to be the subject of study for the current session. Corpora (or, more likely, restricted corpora) may also be deleted completely from the database when they are no longer required.

2.6 Help

Finally, even with an accompanying manual, it is useful to have on-line help facilities. At the bottom of the screen in both the tree viewer and the exploration scheme editor there is a short-list of key-commands to choose from, and a full help screen can be obtained by typing a question-mark.

What more could a corpus linguist want?

2.7 Collocation and concordancing

A number of other features might well have been included in the LDB, but have not as yet. Even if they already exist in other tools, for the sake of fullness, the ability to produce word frequency lists, word+wordtag frequency lists and collocation lists of various types in the LDB would be helpful. Concordancing could be simulated using the exploration scheme on individual words, but hardly in the most elegant fashion, to which users are accustomed elsewhere.

2.8 Probabilistic grammar extraction

One of the key uses of parsed corpora which has not been addressed in the LDB is the automatic extraction of probabilistic grammars. Rather like word lists can be extracted from raw corpora, large sets of context-free

grammar rules can be extracted, along with their frequencies, from parsed corpora. The precise formalism in which the grammar is extracted may vary from context-free rules, to various types of Markov model, or even vertical strip grammars (Atwell *et al* 1991, Souter 1990, Souter and O'Donoghue 1991, O'Donoghue 1991). This omission stems primarily from the fact that parsing work in the TOSCA group at Nijmegen has been non-probabilistic. The grammar used to parse the Nijmegen Corpus is hand crafted rule by rule, and iteratively tested and amended against the corpus. It is formalised in order to be fed into a parser-generator, which contains no knowledge of likelihood of particular structures.[2] Even if the authors have no intention to use the LDB for probabilistic work, any general purpose tool needs to recognise and support the extraction of probabilistic grammars.

2.9 Loading your own corpus

By far the most serious problem with the software (and the book) is its total lack of advice for the proper integration of your own corpus into the database. Despite the authors' observation that the number of parsed corpora generally available was likely to increase rapidly, and the use of a fragment of the Lancaster-Oslo/Bergen (LOB) Corpus as an illustration, the remainder of the text focuses only on the use of the Nijmegen Corpus.

The number of parsed corpora (some of which have been generally available for some time) has indeed risen, to at least half a dozen: The Nijmegen Corpus; the Lancaster-Leeds Treebank (Sampson 1987); the Lancaster/IBM Parsed Corpus (Leech and Garside 1991); the Polytechnic of Wales (POW) Corpus (Fawcett and Perkins 1980, Souter 1989); the Gothenburg/Susanne Corpus (Ellegård 1978, Sampson 1991); the ACL/DCI Penn Treebank. Each of these contains analyses according to different grammars and different notations for representing the tree structure. Figure 1 contains example trees from a few of these corpora.

Nijmegen Corpus (numerical LDB form):

0800131 AT 9102 THIS 2103 MOMENT, 3101 WE 5301 'VE F201 BEEN F801 JOINED A801
0800131 BY 9102 MILLIONS 7803 OF 9104 PEOPLE 3103 ACROSS 9104
0800201 EUROPE, 3501 THIS 2104 ER 1104 COVERAGE 3103 BEING F903 TAKEN A803
0800201 BY 9104 QUITE 2805 A 2505 NUMBER 3105 OF 9106 EUROPEAN 4107
0800202 COUNTRIES 3202 AND 6102 ALSO 8103 BEING F903 TAKEN A803 IN 9104 THE 2105
0800202 UNITED 9906 STATES. 3600 [[9400

Lancaster-Leeds Treebank:

```
A01 68 001
[S[Nns[NPT[ Mr ]NPT][NP[ James ]NP][NP[ Callaghan ]NP][,[ , ],][Ns[NN$[
labour' s ]NNS][JJ[ colonial ]JJ][NN[ spokesman ]NN]Ns][,[ , ],]Nns][V[VBD[
said ]VBD]V] [Fn[Nns[NPT[ Sir ]NPT][NP[ Roy ]NP]Nns][V[HVD[ had
]HVD]V][Ns[ATI[ no ]ATI][NN[ right ]NN][Ti[Vi[TO[ to ]TO][VB[ delay
]VB]Vi][Ns[NN[ progress ]NN]Ns][P[IN[ in ]IN][Np[ATI[ the ]ATI][NNS[ talks
]NNS]Np]P][P[IN[ by ]IN][Tg[Vg[VBG[ refusing ] VBG]Vg][Ti[Vi[TO[ to ]TO][VB[
sit ]VB]Vi][P[IN[ round ]IN][Ns[ATI[ the ]ATI][NN[ conference ]NN][NN[ table
]NN]Ns]P]Ti]Tg]P]Ti]Ns]Fn][.[ . ].]S]
```

Lancaster/IBM Parsed Corpus (Spoken English Corpus Treebank):

```
SK01 3 v
[Nr Every_AT1 three_MC months_NNT2 Nr] ._, [ here_RL [P on_II [N Radio_NN1
4_MC N]P]] ._, [N I_PPIS1 N][V present_VV0 [N a_AT1 programme_NN1 [Fn
called_VVN [N Workforce_NP1 N]Fn]N]V] ._.
```

Polytechnic of Wales (POW) Corpus:

```
189
Z 1 CL FR RIGHT 1 CL 2 C PGP 3 P IN-THE-MIDDLE-OF 3 CV 4 NGP 5 DD THE 5 H TOWN
4 NGP 6 & OR 6 DD THE 6 MOTH NGP H COUNCIL 6 H ESTATE 2 S NGP HP WE 2 OM 'LL 2
M PUT 2 C NGP DD THAT 2 C QQGP AX THERE 1 CL 7 & AND 7 S NGP H WE 7 OM 'LL 7 M
PUT 7 C NGP 8 DQ SOME 8 H TREES 7 C QQGP AX THERE
```

Parsed LDOCE meaning description (Vossen, personal communication):

```
ENTRY{SUBF{FOBZ-} BXC{USG{....} SC{N} RC{.....} } TNR{086828} EW{dairy}
HSNR{00.03} POS{n} GC{-.-} NMD{NP{DE{$D0{a} } RE{HEAD{$N0{shop} } }
POM{CLS(loc){PRDN(pas){SATLT(cl,loc){$cl{where} }
ARG(scnd){NP(co){RE{KE(co){HEAD{$N0{milk} } $ca{'} HEAD{$N0{butter} }
$ca{'}
HEAD{$N0{cheese} } $ca{'} $C0{and} $M0{sometimes} HEAD{$NS{eggs} } } }
$C0{and} DE{$D0{other} } RE{KE(cx){CC{$N0{food} }
HEAD{$NS{products} } } } } }
PRED(cx){Vpass{$BA{are} PREDICATE(past){$VD{sold} } } } } } } } } }
```

Figure 1: Examples of trees from different parsed corpora.

Any general purpose database tool for these corpora needs to be flexible enough to handle the style of their individual notations, and not introduce any restrictions which would render the tool useless for a particular corpus. It is perhaps understandable that the authors did not attempt to tackle this issue, and second-guess the way in which annotators would decide to represent trees in the range of parsed corpora which have recently been developed. Van Halteren (personal communication) explains that the choice to leave out any information on loading your own corpus was a strategic one, because that was seen as the job of an expert database manager in collaboration with Nijmegen:

Then to getting other data into the system. This is not described in the book. This was a design decision. The book is meant for end-users. It is aimed at describing how you can do something with data in the database. Creating new data (and installing new terminal types, etc.) is not supposed to be end-user work, but database manager work. There ought to be a separate book: the LDB manager manual. You may have noticed that it doesn't exist yet.

In fact, the LDB has been used by the LINKS team in Amsterdam to store their parsed version of the meaning descriptions in the *Longman Dictionary of Contemporary English* (Vossen 1991b), reportedly with very little integration effort. Since the publication of the book, renewed interest has fostered new demand for documentation on integrating your own data, and perhaps with a little foresight, an extra section could have been included to cover this area. The LDB requires the following corpus format (van Halteren, personal communication):

1. It must consist of trees.
2. Maximum number of nodes per tree probably (absolute guesswork) about 1000.
3. Each node has two primary label slots (called function and category, but nobody does a semantic check on what you put there) and up to 255 secondary label slots (called attributes).
4. Each of the slots can be filled with strings from a finite set (i.e. three sets: functions, categories and attributes) which the producer of the data can (and must) specify (i.e. you choose your own labels).
5. Slots may be left empty (actually filled with the empty label); e.g. the analyses of the Nijmegen Corpus have no attributes.
6. At the leaves of the trees there are pointers into a separate text file.
7. As no check is made that the leaves of the tree are in the same order as the words in the text file, it is possible to represent discontinuity; it is even possible to search for this.
8. It is also possible to represent ambiguity.

Further details on how to format and load your own data are to appear soon in a short LDB manager manual. Only after a number of different parsed corpora have been tried for loading into the LDB will we be able to say that the database is sufficiently flexible to be a general purpose tool. No standards exist as yet for the format of parsed corpora, but preliminary recommendations are being made by the ACH–ACL–ALLC Text Encoding Initiative (Sperberg-McQueen and Burnard 1990). As it stands, the LDB has yet to prove that it is a tool for the exploitation of all, or even most, syntactically analysed corpora.

3. *Conclusions*

As with most facilities, once they exist, it is easy to take them for granted and only look at their disadvantages, rather than the substantial effort and achievement that has gone into their development. This has undoubtedly been the case with this review of the LDB, which offers computational linguists a sophisticated tool for the handling of a parsed corpus. The book provides necessary guidance on learning to use the software tool. I have attempted to point out areas where improvements could be made in future versions, in the light of the rapid development of corpus-based research in linguistics and natural language processing. I have not discussed factors such as speed, which varies among machines, or program size, which is not prohibitory. Instead I have focussed on the functions which the software supports. The most important worry left in my mind is whether a single tool can possibly cope with the variety of formats in use for parsed corpora. It is to be hoped that through the surveying and standardising of such formats, some sort of norm will be converged upon. Hence a stationary target will be provided, at which developers of software tools such as the LDB can aim.

Notes

1. I refer to Hans van Halteren and Theo van den Heuvel as 'the authors', not for lack of familiarity but for lack of space. I would like to thank Hans van Halteren for his kind help in supplying a copy of the full LDB software and answering many of the questions I have had about its use.
2. In fact, the way the Nijmegen Corpus was parsed is slightly more complex (Keulen 1986). It was initially hand tagged with word classes, and with markers to show the sentence and constituent boundaries, following a case-law manual (CCPP Workgroup 1978). Then, a context-free grammar based on *A Grammar of Contemporary English* (Quirk *et al* 1972) was used to automatically assign full parse trees to the tagged and skeletally structured corpus. The grammar was incrementally improved by test parsing of the corpus.

References

Atwell, Eric, Tim O'Donoghue, and Clive Souter. 1991. *Training parsers with parsed corpora.* Research Report 91.20, School of Computer Studies, University of Leeds.

CCPP Workgroup. 1978. *Manual for coders; A proposal for the syntactic description of English corpus data.* University of Nijmegen.

Ellegård A. 1978. *The syntactic structure of English texts.* Gothenburg Studies in English 43. Gothenburg: Acta Universitatis Gothoburgensis.

Fawcett, Robin, and Michael Perkins. 1980. *Child language transcripts 6–12.* Department of Business and Communication Studies, Polytechnic of Wales.

Keulen, Françoise. 1986. The Dutch Computer Corpus Pilot Project. In *Corpus linguistics II*, ed. by Jan Aarts and Willem Meijs. 127–161. Amsterdam: Rodopi Press.

Leech, Geoffrey, and Roger Garside. 1991. Running a grammar factory: The production of syntactically analysed corpora or 'treebanks'. In *English computer corpora: Selected papers and research guide*, ed. by Stig Johansson and Anna-Brita Stenström. 15–32. Berlin: Mouton de Gruyter.

O'Donoghue, Tim. 1991. *The vertical strip parser: A lazy approach to parsing.* Research Report 91.15, School of Computer Studies, University of Leeds.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik. 1972. *A grammar of contemporary English.* London: Longman.

Sampson, Geoffrey. 1987. The grammatical database and parsing scheme. *The computational analysis of English: A corpus-based approach*, ed. by Roger Garside, Geoffrey Leech and Geoffrey Sampson. 82–96. London: Longman.

Sampson, Geoffrey. 1991. Analysed corpora of English: A consumer guide. In *Computers in applied linguistics*, ed. by Martha Pennington and Vance Stevens. Multilingual Matters.

Souter, Clive. 1989. *A short handbook to the Polytechnic of Wales Corpus.* Bergen: Norwegian Computing Centre for the Humanities.

Souter, Clive. 1990. Systemic functional grammars and corpora. In *Theory and practice in corpus linguistics*, ed. by Jan Aarts and Willem Meijs. 179–211. Amsterdam: Rodopi.

Souter, Clive and Tim O'Donoghue. 1991. Probabilistic parsing in the COMMUNAL Project. In *English computer corpora: Selected papers and research guide*, ed. by Stig Johansson and Anna-Brita Stenström. 33–48. Berlin: Mouton de Gruyter.

Sperberg-McQueen, C.M., and Lou Burnard. 1990. *Guidelines for the*

encoding and interchange of machine-readable texts, TEI P1. Chicago and Oxford: ACH–ACL–ALLC.

Vossen, Piek. 1991. Polysemy and vagueness of meaning descriptions in the Longman Dictionary of Contemporary English. In *English computer corpora: Selected papers and research guide*, ed. by Stig Johansson and Anna-Brita Stenström, 105–123. Berlin: Mouton de Gruyter.

Rejoinder by Hans van Halteren

It is always hard on an author to have to read a reviewer's opinion of his work and not to be able to react. I am therefore very grateful to the editors that they have enabled me to respond to the review above.

Let me start by saying that on the whole this particular review is no reason for negative feelings on my part. In fact, I agree with most of the points it raises. For example, I, too, think that a mouse-based user interface is preferable to a keyboard-based one. As to why it has not been implemented, then, I can only offer an excuse: when a software system reaches a certain complexity, it becomes more efficient to create a completely new version than to try and add on major new components. A mouse-based user interface has such extensive implications for the system as a whole that indeed it will have to wait for the next version.

There are, however, some points on which I do not agree. First I would like to address Section 2.2. I do not think that the activities can be replaced by a mechanism which just sends information to the screen. The activities provide much more variability and control than that, e.g. it is possible, although the reviewer doubts this (Section 2.8), to extract a grammar from a parsed corpus (see exercise 4.12.2 for an indication of how to do this). I do think that a simplified, and more user-friendly, mechanism could be added as an alternative for casual users. Again, such an addition is planned for a later version.

Secondly, and this is the main reason to respond in this way, I totally disagree with Section 2.9. The doubts raised here are probably based on a misconception about the Nijmegen Corpus/CCPP analysis and the LDB format. The examples in the book do not exclusively use the Nijmegen Corpus (although probably all screen dumps of trees are from this corpus). Also, the example from the Nijmegen Corpus in the review

does not show the numerical LDB format, but the horizontal CCPP tagging format. The LDB uses no specific one-dimensional representation, but stores the analyses as actual tree structures. As a result, none of the corpora presented pose any problems regarding storage and access through the LDB. All of them have the required structure, viz. trees with simply labelled nodes. They only differ in the way this structure is represented. A simple transformation, costing no more than a couple of days to effect, suffices to import such data. This time estimate is of course only valid when the work is done by an experienced database manager. That is why we suggest that anybody who wants to import data contact us for instruction. Only after we have gained more experience with the diversity of treebank notations, will we try to put these instructions in writing as part of an LDB manager manual.

Reviews

Karin Aijmer and Bengt Altenberg (eds.). *English corpus linguistics: Studies in honour of Jan Svartvik.* London and New York: Longman, 1991. xi + 338 pp. ISBN 0–582–05931–3 cased; 0–582–05930–5 paperback. Reviewed by **Geoffrey Sampson**, University of Sussex.

This book presents itself as a *Festschrift* for Jan Svartvik and a survey of the state of the art in English-language corpus linguistics, which is one of Svartvik's chief areas of activity (though not the only one). After a brief introduction by the editors which is essentially an annotated contents list, it contains nineteen chapters by authors from many parts of the world, namely: Geoffrey Leech; M.A.K. Halliday; Jan Aarts; Wallace Chafe and collaborators; Sidney Greenbaum; Graeme Kennedy; Göran Kjellmer; Antoinette Renouf and John Sinclair; Peter Collins; Charles Meyer; Dieter Mindt; Gabriele Stein and Sir Randolph Quirk; Douglas Biber and Edward Finegan; David Crystal; Anna-Brita Stenström; Gunnel Tottie; Matti Rissanen; Ossi Ihalainen; Stig Johansson.

I find the book disappointing, in part because it falls between two stools, being neither a *Festschrift* in the normal sense nor a systematic topic survey.

As a *Festschrift* it is rather odd. The subtitle 'Studies in honour of Jan Svartvik' appears on the title page, there is a dedication, and a half page in the prelims repeats that the book is intended as a tribute to Svartvik, saying that his contributions 'are well known and need no special presentation'; and that is far as the Svartvik theme goes. A reader who is less familiar than the editors with Svartvik's career will learn little about it here; an Appendix, which might have been used for a Svartvik bibliography or the like, is instead devoted to a listing of available English corpora. Many of the contributors are Svartvik's pupils, or fellow-members with him of the Quirk quadrumvirate of English grammarians, but others have no particular Svartvik connection that I can detect. (Yet the contributors do not constitute a roll-call of 'all the great names of English corpus linguistics', either: the men who started it all, Nelson Francis and Henry Kučera, continue to flourish, but neither appears here.) The book does not even identify what special career milestone made 1991 the right time for a Svartvik *Festschrift*. (It happens that I was present when the book was formally handed over: from memory, the occasion was a 'round' birthday.) I believe Svartvik merited less cursory treatment than this.

On the other hand, the book is in no sense a comprehensive, organized

account of the current state of the discipline (if a 'discipline' is what corpus linguistics is). It reads more as if nineteen individuals or partnerships were pressed to offer something or other: some wrote pieces specially for the book, others sent in what they happened to have lying around - and one or two perhaps seized the chance to place a paper that had failed to find a home elsewhere (one contributor forgets to update his remark that the LOB texts 'are now twenty-five years old' – in 1991 the correct figure was thirty).

This is not to say that the book is valueless. It is useful, for instance, to have an article by Wallace Chafe and his team on the Corpus of Spoken American English which they are creating at Santa Barbara, which is on the way towards filling the so-far empty slot in the 2 x 2 classification written v. spoken, British v. American. And a number of others chapters summarize and give references to significant bodies of research by the respective authors which are published in full in book form or in journals where one might not think of looking for them, so this book offers a convenient way of checking the work.

I do not think, though, that someone who knows about linguistics and/or computing, and who wants to discover what is going on in corpus linguistics and whether he should get involved in it, is likely to find his enthusiasm stirred by this book. Indeed, there is a characteristic running through several of the contributions which an outsider might find positively offputting. I take it for granted that corpora are means to ends: corpus-based research is only likely to be interesting if it grapples with problems that one would have wanted to solve anyway, but which can only be solved, or can be solved more easily, with the use of a corpus. Too many of these pieces read as if their authors had said to themselves 'Now we've got this corpus, what can we do with it?'

This may be a consequence of the format of the book, which encourages brief studies of isolated and therefore in some cases rather trivial topics. But I suspect it may also reflect a real tendency to sterile, inward-looking activity, against which this research community needs to be on its guard. Corpus linguists ought to spend less time talking to corpus linguists, and more time talking to other researchers whose work could be advanced by using corpora.

Stig Johansson and Anna-Brita Stenström (eds.) *English computer corpora: Selected papers and research guide*. Mouton de Gruyter 1991. vii, 402pp. Reviewed by **W. Nelson Francis**, Brown University.

This book is one more in the growing list of collections of studies in computer corpus linguistics resulting from the annual ICAME conferences. It presents a selection of papers, mostly from 1989 conference, in which ICAME celebrated its tenth anniversary by returning to Bergen, where the first conference was held in 1979. Year by year the volumes of collected papers reveal the advance of computer linguistics in range of subject matter, technical complexity, and contribution to linguistic theory.

Following a succinct introduction by Stig Johansson, defining the subject and summarizing the individual contributions to the volume, there are twenty papers arranged under eight headings: Probalistic grammatical analysis, Syntax, Lexis, Speech, Regional/social variation, Specialised corpora, Software, and Reference.

1. Probabilistic grammatical analysis

In a brief paper, Steven J. DeRose reviews various probabilistic methods for performing lexical syntactic tagging of corpora. He finds that as a result of the work on CLAWS and his own Volsunga, 'stochastic tagging is robust under a wide range of algorithmic variations', reaching accuracy between 93% and 94% in both English and Koiné Greek.

The remaining two papers in this section deal with the application of probabilistic methods to automated or semi-automated parsing programs. In 'Running a grammar factory: The production of syntactically analysed corpora or "treebanks"', Geoffrey Leech and Roger Garside describe progress at Lancaster University in developing a 'skeleton parsing' method of parsing natural English rapidly and with satisfactory accuracy. Basically the method is to set up a collected sample of manually parsed sentences – a 'treebank' – from which rules can be derived which will carry out the parsing of raw text. The first effort was the development of the Lancaster-Leeds treebank by Geoffrey Sampson, in which 45,000 words, already tagged by CLAWS, were manually parsed according to a system developed by Sampson. It turned out that the number of rules in even a corpus of this small size was very large, with 'a large proportion' of the rules occurring only once, and the total number increasing 'at a scarcely diminishing rate' as the size of the corpus was increased. It looked as though this type of grammar of English was open-ended. The Lancaster group continued by simplifying the Sampson model. The ultimate result was the 'skeleton treebank'. As a result of

their experience with four successive models, each somewhat simpler than the last, Leech and Garside reached the conclusions 'that (a) tagging is best done automatically with correction by a human post-editor, but (b) skeleton parsing can best be done by the human analyst, with the aid of a fast input program' (29).

In 'Probabilistic parsing in the COMMUNAL project', Clive Souter and Tim F. O'Donoghue briefly describe various parsing procedures such as shift-reduce parsing and chart parsing. They come out in favor of a probabilistic model, the Realistic Annealing parser. They describe and illustrate this, rather too briefly for easy understanding. They conclude that this model 'is still very much at an experimantal stage, where annealing schedule parameters are being tested, and more efficient methods of implementation are being considered' (44). They have not done any testing of speed or accuracy. It remains to be seen how valuable it will be.

2. *Syntax*

This section includes two papers using corpus material in the analysis of specific syntactic problems in English. In 'On the exploration of corpus data by means of problem-oriented tagging', Pieter de Haan deals with postmodifying clauses in the English noun-phrases. He extracts 2,430 such clauses from the Nijmegen Corpus and classifies them according to a numerical coding system which greatly reduces their bulk. This work was all done manually, since the Nijmegen Linguistic Data Base (LDB), which would permit automation, was not yet available (see Van Halteren & van den Heuvel 1990). He then uses the computer to make a series of statistical studies of the pattern, function, position, type, and complexity of the postmodifying clauses and their containing noun phrases. He puts forward his results as tentative, looking forward to greater insight to be derived from a syntactically fully analysed corpus such as the LDB.

Christian Mair raises the question 'Quantitative or qualitative corpus analysis?' and proposes an answer based on complement clauses in the Survey of English Usage (SEU). Basing his work on the SEU before it had been computerised, he gives examples of complex infinitival complements characteristic of both written and spoken discourse. He concludes: 'The role of the corpus, after all, is not only to provide a limited and representative data-base for statistical analysis, but also to provide authentic and realistic data, the close reading of which will

allow the linguist to approach grammar from a functional and discourse perspective' (77).

3. *Lexis*

This section consists of two papers on lexical ambiguity and polysemy. In 'Automatic parsing meets the wall', Magnar Brekke deals with the treatment of the noun wall by several dictionaries, some monolingual and others bilingual Norwegian-English. He finds that they agree on recognizing five 'referential distinctions' for *wall*, two of them primary and the others more or less figurative. He concentrates on the first two: 'vertical side of room or building' and 'long, narrow, vertical dividing structure'. These he calls the 'house related' sense and the 'garden related' sense. Using the microfiche concordances of Brown and LOB, he extracted all uses of *wall* in these two senses, a total of 140 in Brown and 126 in LOB. He found that even the short span of a KWIC concordance supplied enough context to clearly disambiguate over 50% of the cases. He goes on to analyse the examples in semantic detail, with the ultimate aim of developing objective rules to facilitate disambiguation by computer. His choice of *wall* is particularly apt in regard to machine translation, since both Norwegian and German have seperate words for the two major senses of English *wall*.

Pick Vossen writes about 'Polysemy and vagueness of meaning description in the *Longman dictionary of contemporary English*'. He analyses a number of entries from LDOCE, distinguishing basic from extended meanings and diagramming the semantic relationships among them. Studies like this should be of great value to lexicographers facing the universal lexical problem of polysemy and how to relate the various meanings of a word in a way that accords with a system of hyponymy which is intuitively understood by the native speaker.

4. *Speech*

Two papers in this section deal with specific problems in the transcription and prosodic analysis of intonational features of natural spoken English. Using the Lancaster/IBM Spoken English Corpus (SEC), Gerry Knowles deals with the vexed problem of tone group boundaries and their relation to syntactic constituents. He cannot be said to have shed much light on the problem. In fact, he concludes by denying the existence of tone group boundaries entirely. Instead, 'Discontinuities must therefore be identified and labeled in their own account, and not bound to the kind

of tone group theory that underlies convential transcription' (160).

Anne Wichmann presents 'A study of up-arrows in the Lancaster/IBM Spoken English Corpus'. The term 'up-arrows' derives from the symbol used in transcription to denote an 'upward pitch excursion' to 'reset' the natural downward trend of pitch through an utterance designated by Ladd and others as 'declination'. Wichmann attempts to classify these by position and by function. The fact that she frequently mentions phenomena which 'have not yet been examined' or need further research, reveals that this is a report of work in progress rather than accomplishment. But it does show the importance of carefully transcribed corpora of natural spoken English as data for an understanding of speech and possible text-to-speech computer programs.

Bengt Altenberg continues his study of recurrent word combinations in the London-Lund Corpus of Spoken English in 'Amplifier collocations in spoken English'. He here deals with stock phrases used to intensity the meaning of a key adjective, adverb, or verb. Following Quirk et al. 1985, he divides amplifiers into two sets, maximizers and boosters. Maximizers 'denote an absolute degree of intensity and therefore occupy the extreme upper end of the scale', while boosters 'denote a high degree but without reaching the extreme end of the scale' (128). He goes on to describe and discuss the collocations and frequencis of 29 maximizers ranging from *quite* and *utterly*, and 69 boosters, ranging from *very* to *incredibly*. He presents tables of combinations of maximizers plus adjective/adverb, the most frequent being *quite sure*, and booster combinations, with *very* leading the list, though yielding to others in special phrases like *terribly helpful* and *bloody cold*. He concludes by pointing out the value of the corpus for such studies: 'it can serve to enrich existing description in grammars and dictionaries, provide a basis for comparisons with other varieties, and highlight areas where supplementay corpus or elicitation studies are needed' (145).

5. *Regional/Social variation*

The original large corpora of English – SEU, Brown, and LOB – present a body of English which can be classed as standard. The only significant variation they display is stylistic, as revealed by the various genres sampled, and, in the case of SEU, the contrast between writen and spoken language. More recent corpora of other varieties have opened up the possibility of regional variation (this, of course, was revealed in the comparison of Brown and LOB; see Hofland and Johansson 1982). Two papers in this section deal with regional differences as

shown in another regional native variety, Australian, and the only non-native variety as yet collected, Indian English. In '*Will* and *shall* in Australian English', Peter Collins, one of the collectors of the Australian corpus, compares the usage of these two modals in Australian with British and American, as shown in LOB and Brown. He finds several interesting contrasts, perhaps the most being *shall*, which 'appears to be almost obsolete in Australian English' (184).

Gerhard Leitner deals with the only non-native corpus following the Brown/LOB tradition, the Kolhapur Corpus of Indian English. He discusses the vexed question of the status of such varieties of English: are they on a par with the post-colonial native varieties – American, Canadian, Australian, and New Zealand – or are they to be considered merely as deviant, suffering from adstratal lexical variation and 'mistakes' of grammar? Leitner inclines to the former position, which has been most ardently defended by Braj Kachru (most recently in Kachru 1990). He discusses differences in lexico-syntactic matters, which he finds 'appear not to be of a systemic, but of a quantative nature' (228). On the other hand he finds new semantic and pragmatic distinctions occurring which indicate more fundamental departure from the British norm. He questions whether the close adherence to the Brown/LOB format which Shastri adopted in Kolhapur is the best way to present the intrinsic nature of the non-native corpus, representing as it does a culture contrasting in many ways with that of the U.S. and U.K. He ends by looking forward to the forthcoming International Corpus of English which will sample at least seventeen varieties of English world-wide. (Greenbaum 1991).

Yet another form of varietal difference is that between the standard language and regional or social dialects. The principal effort to study this is the Helsinki Corpus of Modern English Dialects being collected by Ossi Ihalainen. Its aim is to collect rather large samples of continuous discourse in rural dialects in England, to permit extension of the primarily lexical-phonological *Survey of English Dialects* into syntax and discourse. In his paper in this volume, Ihalainen compares the grammatical subject in a corpus of Somerset dialect with the London-Lund Corpus of Spoken [Standard] English. He deals with types of subject, ellipsis of subject, omission of existensial *there*, and the usage of impersonal *they* in the Somerset dialect, all of which are noticeably different from standard usage. This is an innovative field of corpus study which shows great future possibilities.

6. *Specialised corpora*

Most of the corpora recently collected have been broad in scope, attempting to cover a wide variety of subject matter, as in the fifteen genres of prose included in Brown, LOB, and Kolhapur. But there is need also of specialised corpora, covering single fields more intensively. One which is described by Karen Lauridsen and Dorrit Faber is the Danish-English-French corpus in contract law being compiled at the Copenhagen and Aarhus schools of business. This is the only trilingual corpus so far projected, and one limited to a rather narrow subject field. Lauridsen and Faber describe the format of the separate corpora, classified by type of material and by legal theme. They have been careful to select the individual samples according to a preconceived weighting scale. This has not been easy, considering the basic difference between English common law and French civil law. The analysis now going on at Aarhus and Copenhagen will determine how useful this corpus will be.

Another type of corpus is dealt with by Magnus Ljung in 'Swedish TEFL meets reality'. This is the body of reading material in English used in teaching English in Swedish gymnasia. Ljung compiled a corpus of approximately 1.5 million words. He compares high frequency items in his GYM corpus with that used for the COBUILD dictionary. He finds in general that in the GYM texts 'There is heavy emphasis on concrete and uncomplicated matters and a dearth of abstractions and words relating to the organization of society. There are also indications that the texts included tend to be of fairly simple, narrative kind' (255). He feels this is neglecting the major uses which advanced students of English will make of the language after they leave school, such as 'reading newspapers, reading and producing reports and manuals and following newscasts on the media' (255). It seems to him that the reading in the advanced years of English study should prepare the students for these activities.

7. *Software*[1]

In the first of the three papers on corpus-related software, Benny Brodda discusses his PC Beta program, 'a PC oriented tool for corpus work in the broadest possible sense' (259). After a brief description of the principles underlying the system, he presents – in order of increasing complexity – a number of specific applications of the program to corpus research. These applications include text normalization and 'investigation'

(relevant to the development of corpora from diverse source texts), excerption and concordance generation. In the course of these examples, a point is made concerning the intentional limitations of PC Beta; since computer systems normally include utility packages for such functions as sorting files and counting tokens, these functions are not provided in PC Beta. Rather, PC Beta creates files (by use of its pattern-matching capability) that then can be handed over to the appropriate packaged routines for further processing.

A specific example is then provided of PC Beta's application to a lingusistic problem: the excerption of passive structures from an unannotated corpus. An extended discussion is provided as well of excerption evaluation, along the information-retrieval dimensions of 'precision' (i.e. percentage of retrieved material that is relevant to the task at hand) and 'recall' (i.e. percentage of the material desired that is actually retrieved). Noting that 'the key to successful corpus investigations is interactive work' (277). Brodda sketches a procedure for such investigations that involves fine-tuning the rules defining the retrieval in successively larger spans of text.

An intriguing extension to the system is then considered; once it has been finetuned to excerpt a given structure with acceptable degrees of precision and recall, it can then be used to identify this structure for the purpose of (non-recursive) surface-structure parsing. A study along these lines for Swedish is reported, with promising results. Brodda notes that these results may depend on language-specific features, but the approach in general seems worthy of further investigation.

In recent years, advances in microcomputer technology have made it possible to perform the same types of corpus text-processing (e.g. production of concordances and indexes) that were once confined to large mainframe computers. In his paper, Knut Hofland examines five such programs that are generally available for these purpose: WordCruncher, MicroOCP, CLAN, TACT, and Free Text browser (current sources of distribution for each program are also supplied).

Considerable attention is devoted to WordCruncher, which was originally developed at Brigham Young University, and is now distributed by Electronic Text Corporation. It consists of two components: IndexETC (which indexes a given text) and ViewETC (which operates on pre-indexed texts to look up references or generate concordances). A discussion of the functions and capabilities of both programs is provided, along with sample screens; in conclusion, Hofland finds that WordCruncher is simple to install and use, and is a 'powerful program for swift searches in large texts' (291), though it has various minor weaknesses; e.g. a lack

of flexibility in its reference system, only one possible sort order, and memory limitations.

Next, Hofland examines MicroOCP, a microcomputer implementation based on the mainframe concordance program OCP, developed at Oxford. Like its mainframe counterpart (with which its files are fully compatible), MicroOCP is a 'batch program for the production of word lists, indexes and concordances' (291). A presentation of various components of the system is made, again including sample screens from its menu-driven user interface. Hofland concludes that the program is very powerful and flexible, easy to install and supplied with a good tutorial. Its main weakness is its 'severe demand on computing time' (295), though (as is examined further below), this depends on the machine configuration on which the program is run.

Next, Hofland briefly mentions CLAN (a system of text-analysis programs that was originally developed at Carnegie-Mellon University for the analysis of child-language databases, but now also expandable to other texts) and concludes that the programs are mainly 'useful for those who need facilities for making simple concordances, word lists, or for studying patterns in texts' (296).

The final two programs examined in this paper are freeware programs: TACT (developed at the University of Toronto by John Bradley and Lidio Presutti) and Free Text browser (a Macintosh program in Hypercard developed by Mark Zimmerman). Hofland examines both of these programs in detail similar to his treatment of WordCruncher and MicroOCP, along with sample menu screens. He finds TACT to be similar to WordCruncher, but with considerably faster indexing capabilities, and concludes that it is 'well done' and has 'qualities not found in commercial programs' (301). The Free Text browser receives similar high marks, though he notes that (despite fast indexing capabilities) it lacks some options for corpus work. In addition (unlike the other programs examined) it is tailored for a specific machine configuration: Macintosh/Hypercard. A further point concerning both freeware programs is the accessibility of their authors and their willingness to consider suggestions for revisions and improvements to their programs.

After his examination of each of the five programs, Hofland performs a comparative test of their capabilities by running them on the same task (generating a concordance of a 155K text) on a variety of processors (with the exception of Free Text, which is confined to a Macintosh – on which, however, it was considerably faster than any of the other programs). A table of results is presented and analyzed, and reveals the importance of taking into account the machine configuration (i.e. pro-

cessor, type and speed of peripherals, etc.) on which a given program is run, since the results can vary by as much as an order of magnitude.

In his concluding remarks, Hofland notes that there are other programs available for indexing and concordancing, and suggests further reading – nevertheless, his presentation gives a valuable introduction to a variety of important programs and standards of comparison.

In the brief final paper of this section, Jacques Noël presents a Unix/Awk approach to searches in dictionaries and corpora consisting of multiline records. 'Awk' is a pattern-matching language, created for use with Unix operating systems, which was designed to be a 'convenient and expressive programming language that can be applied to a wide variety of computing and data-manipulation tasks' (Aho, Kernighan, and Weinberger 1988:1). Its advantage in the application to corpus research lies in the fact that the varying-length multiline records up to approximately 3,000 characters long 'behave as single lines under Unix/Awk' (307). This allows the segmentation of the input texts into multiline records 'consisting of natural or logical units' (307) (a list is given of the dictionaries and corpora in the Liège archive that have been converted to this format, totalling well over 100 megabytes). Noël notes that this approach represents an 'economical compromise solution' between the operation of two of the commercial programs used at Liège (both of which are also examined in Hofland's paper): CLAN (which does not require specially prepared source files) and WordCruncher (which, though 'must useful in our day-to-day work' (307), has high storage requirements).

In addition to outlining the steps necessary to convert a database into the multiline record Awk format, Noël presents some examples of Awk retrievals, as well as a brief discussion of the pros and cons of using this approach as opposed to, for example, WordCruncher. He concludes that the latter is 'ideal for complex searches with immediate feedback and browsing in large files' (311), but notes that Awk databases (unlike those of WordCruncher and other convential indexing programs) have the advantage of not requiring re-indexing each time a change (however slight) has been made to them.

8. Reference

The reference section which concludes the volume consists of two very valuable compilations. A group from Lancaster – Geoffrey Leech, Lita Taylor, and Steven Fligelstone – present a list of 36 English corpora, giving details of content, format, and availability. Bengt Altenberg

continues his comprehensive bibliographical work with a 42-page bibliography of publications relating to English computer corpora.

In sum it may be said that this book is required reading for both those actively carrying on computer corpus research and also for those attempting to keep up with this rapidly moving field. The editors are to be congratulated on a judiciously selected and carefully prepared volume.

Note

1 The papers in this section are reviewed by Andrew W. Mackie.

References

Aho, Alfred V., Brian Kernighan, and Peter J. Weinberger 1988. *The AWK programming language*. Reading, Mass.: Addison-Wesley.

Greenbaum, Sidney 1991. ICE: the International Corpus of English. *English Today*, NO. 28, 3-7.

Hofland, Knut and Stig Johansson 1982. *Word frequencies in British and American English*. Bergen: Norwegian Computing Centre for the Humanities.

Kachru, Braj B. 1990. *The alchemy of English: The spread, functions, and models of non-native Englishes*. Urbana and Chicago: University of Illinois Press.

Quirk, Randolph, Sidney Greenbaum, Geoffrey Leech and Jan Svartvik 1985. *A comprehensive grammar of the English language*. London: Longman.

Van Halteren, Hans and Theo van den Heuvel 1990. *Linguistic exploitation of syntactic databases: The use of the Nijmegen Database program. Linguistic Database program*. Amsterdam: Rodopi.

Merja Kytö. *Variation and diachrony, with Early American English in focus. Studies on CAN/MAY and SHALL/WILL*. Bamberger Beiträge zur Englischen Sprachwissenschaft / University of Bamberg Studies in English Linguistics, 28. Frankfurt am Main, etc.: Peter Lang, 1991. Pp xv, 420. Reviewed by **Udo Fries**, University of Zürich.

This book consists of three unequal parts. Part I, called 'Frame of Reference', is an original study covering 79 pages, which presents the

background for the much longer second part, 'Studies' (273 pages), which is, in fact, a reprint of six papers published or to be published between 1986 and 1992. This is followed by Part III: a 63-page up-dated bibliography for parts I and II. The odd fact that each of the six papers presented in part II appears with a bibliography of its own is due to the fact that 'the regulations concerning the format of a cumulative dissertation at the University of Helsinki do not allow revisions in the individual papers' (p. xv). Surely, this is a regulation which needs reconsideration by the University authorities: it is detrimental to the advancement of learning when authors are prevented from making changes they might feel necessary – and it is detrimental to the environment, since it results in unnecessary pages of printed paper that nobody wants to have: in our case, about 32 pages of footnotes and individual bibliographies, plus any number of repetitions in the introductions to the individual texts.

In Part I Kytö sets out to present the aims and methods of her study, to survey Early American English, to introduce the reader to her Corpus of Early American English, and, finally, to discuss variation analysis with respect to the English modals. Kytö aims at a diachronic variational study of the English modals, in particular of *can* (*could*) and *may* (*might*), and to a lesser extent of *shall* (*should*) and *will* (*would*). Four of the six chapters (papers) of Part II deal with these modals in early American English, and one chapter (the first of two on *shall* and *will*) is about the change from Middle English to Early Modern English.

Kytö argues that early American English and the development of the modals form an ideal combination for a variational study. It is a period crucial to the development of the modal system and it has been largely neglected by scholars as an area for socio-historical linguistics. She goes on to define and describe the area of her study (New England), and the period it covers (five decades before and five after 1670). She treats problems of comparison between British and American English, discusses stages of an emerging variety, and factors influencing language change, always with New England in mind: the problem of 'colonial lag', geographical and social mobility, and the importance of education and social networks (the role of families in early American society). All this is presented in a lucid, though not very detailed manner, but we do find references to all the important sources, albeit occasionally only in a footnote referring us to yet another footnote in Part II (e.g. there is no discussion of Görlach's views of 'the myth' of colonial lag, referred to in footnote 20 of Chapter 6, for which Part I would have been the ideal place).

By way of introduction to her corpus of early American English, Kytö discusses problems connected with the relation of written texts to the spoken language, existing genres of the time, questions of authorship, the origin of and possible changes in manuscripts, textual parameters and informant properties (age, sex, generation). Again, there is no space for a detailed discussion of the more general problems: the difference for example, between genres, traditional text types (appropriately called *Textsorten* in German, rendered on one occasion [p. 41] as 'text sorts' by Kytö) and text types proper. The texts included in the corpus are arranged according to genre/author and period, and in a second table speech-based vs. non-speech-based texts. The whole corpus consists of roughly 1.1 million words.

In section 4 of Part I, Kytö at last turns to the analysis of modals, and immediately shows that she is fully at home in her topic. This is a very lucid survey of approaches (up to 1990) to modal auxiliaries, syntactic and semantic changes, the difference between more traditional approaches and diachronic variation analysis.

Part II contains the six papers that constitute the major contribution of this volume. The first gives convincing statistical evidence of the emergence of epistemic *might* in colloquial early American English. The second chapter widens the perspective considerably by including *can* and *could* in the discussion. Here, Kytö discusses the largest group of examples: the non-epistemic, non-past use of the four modals, carefully distinguishing between speech-based and non-speech-based texts, formal and informal use, and occurrence in affirmative and negative sentences.

Chapter Three is mainly concerned with the modals in Old and Middle English, based on parts of the Helsinki Corpus. This is a very extensive and detailed study, which brings in a number of points not raised so far. One additional problem is the occurrence and gradual disappearance of *can* and *may* as full verbs in Old and Middle English. Furthermore, Kytö discusses the trend from root to epistemic modality, and she includes, for the non-epistemic instances, notional sub-categories (ability, permission, neutral possibility), negation, and type of subject (animate, inanimate).

The final chapter on *may/might* and *can/could* extends the analysis in yet another direction by comparing the findings in the Early American Corpus with findings in early British English. The Helsinki Corpus for Early Modern (British) English includes texts from 1500 to 1710; the American corpus, as we have seen, from 1620 to 1720. This is the most detailed part of the study, which is justified by its importance for the study of the new emerging variety of American English. It sheds

new light on the phenomenon of 'colonial lag' in particular in written texts.

The section on *will/would* and *shall/should* in Part II begins with a short paper on these words in Middle and Early Modern English (1150-1710), based on the Helsinki Corpus. Kytö is interested in the rise of *will* which reaches a first peak beetween 1570 and 1640, after which the influence of grammarians, who oppose the use of *will* for the 1st person, can be felt. This is well known, but Kytö manages to give new and convincing figures distinguishing not only between the persons (1st, 2nd, 3rd person), but also between the text types involved: handbooks, sermons, trials, fiction and many others.

The long final chapter (pp. 277-344) relates early British to early American English use of *shall* and *will*. Kytö studies the influence of text type, level of formality, tense and aspect, clause types, active and passive constructions, animate and inanimate subject, dynamic and stative meanings of the main verb, epistemic and root uses, and the sex and participant relationship of the author. As she points out in the final summary 'conservatism (rather than innovation) characterizes the development of modal auxiliaries in early American English' (p. 353).

Although Kytö draws conclusions concerning the overall history of the modals in English, she is is fully aware of the limitations of corpus linguistics, and that the reliability of the results depends on the size and quality of the corpus. A quantitative study depends, of course, on figures: these are presented in about 100 tables. The number of quotations in all the six articles is only about 120. Here, the reader might have preferred somewhat greater generosity – which would surely have been forthcoming if it had not been for the Helsinki University regulations. Even so, this is a book well worth reading, and it will undoubtedly become a standard reference work on the history of English modals. The collection proves to be more than the sum of its parts: it gives ample evidence of the importance of corpus work, if this is carried out in as careful a manner as it is here, and gives a good idea of the usefulness of the Helsinki Corpus of English Texts. Diachronic and Dialectal, including its sub-corpus[1] of early American English.

Note

1 The reviewer is grateful that Kytö has returned to using the plural form *corpora* and does not go on referring to *corpuses* (cf. p. 84, originally published in 1986).

Ian Lancashire. *The humanities computing yearbook 1989–90: A comprehensive guide to software and other resources.* Oxford: Clarendon Press, 1991. 18 + 701 pp. ISBN 0–19–824253–0. Ian Lancashire (ed.). *Research in humanities computing 1: Selected papers from the ALLC/ACH Conference, Toronto, June 1989.* Oxford: Clarendon Press, 1991. 16 + 353 pp. ISBN 0–19–824251–4. Reviewed by **Stig Johansson**, University of Oslo.

The new yearbook (HCY89–90) is a follow-up of a similar yearbook published in 1988 and edited by Ian Lancashire and Willard McCarty (HCY88). Whereas the latter surveyed 'several decades of research and instructional applications, mainly from the editors' perspective, English studies' (HCY89–90, p. xi), the new yearbook focuses on a more limited period (1985–90), adopts a revised taxonomy, and draws on its Editorial and Advisory Boards, with representatives from a range of disciplines. There are three main parts: Disciplines (pp. 1–380), Methods and tools (pp. 381–518), and Resources (pp. 519–570). These are followed by a list of abbreviated references (pp. 571–555) and a full index (pp. 597–701).

Scholars working in the field of humanities computing will find this reference work invaluable. Some sections are concerned with areas which should be of special interest to readers of this journal: Computational linguistics (pp. 32–66), English language instruction (pp. 75–94), Linguistics (pp. 157–181), including a sub-section on Corpus linguistics (pp. 159–170), and the sub-sections on English in the section on Natural languages and literatures (pp. 232–245).

To show more specifically what sorts of information we find and to test the accuracy and completeness of the information given, we shall briefly examine the treatment of Corpus linguistics. The field is first briefly defined, then there is an annotated bibliography of some forty items, followed by an annotated list of software and data sources. The information seems accurate, with some minor exceptions. The developer of the Gothenburg Corpus is Alvar Ellegård, University of Gothenburg, Sweden (this is not the name given in the book). The reader mistakenly gets the impression that the texts provided with the Linguistic Database are from 'the 1.5-million-word *Nijmegen Corpus* of materials written to be read [sic] after 1975' (p. 166). There are two quite different corpora associated with Nijmegen: a smaller collection called the Nijmegen Corpus (this is the one provided with the Linguistic Database), and a 1.5-million-word collection called the TOSCA Corpus. The reader is also led to believe that the dialectal part of the Helsinki Corpus is

completed. It is still under development. It is in fact often difficult to know whether an item is completed or under development and, if completed, how and on what conditions it is available. The references provided with each corpus or piece of software, however, make it possible for readers to explore the matter further on their own.

The survey of Corpus linguistics includes a sprinkling of references for other languages, but it is interesting to note that it has a heavy English slant (perhaps in part reflecting the role of ICAME?). Nevertheless, the list of references is less comprehensive and the list of corpora less detailed than those found in the reference section in Johansson and Stenström (1991: 319–396).

It cannot really be expected that a general reference work like HCY89–90 should be as detailed in a particular area as a more restricted reference work. Its most important role is to show scholars working in the field how their work relates to humanities computing in general. This is what makes the book so fascinating. We must be grateful to the editor and his many advisers for having provided us with this map over the whole territory of humanities computing, which has become too vast for any one of us to explore fully on our own. It provides a perspective on our own work and opens the field for cross-fertilization.

Needless to say, there are problems of classification, but with the detailed index the reader can easily find the way through this large volume. It could be argued that most of the information given in the yearbook could be more appropriately made available in a continually updated database, but it is convenient to have a printed reference work (perhaps there might be an on-line database in addition?). It is highly desirable that HCY88 and HCY89–90 are followed up in a couple of years' time by a new yearbook. One way of making such a yearbook even more valuable would be to include surveys with more comment (state-of-the-art articles) rather than annotated alphabetical lists. To keep within manageable proportions, it may be necessary to restrict such surveys to areas where there have been particularly significant developments.

The yearbook provides the map; *Research in humanities computing 1* permits a closer look into some particular areas of humanities computing. It contains twenty papers selected from among those presented at the 'Dynamic Text' conference in Toronto in 1989, and is the first in a series of annual publications on the state of the art in humanities computing (edited by Susan Hockey and Nancy Ide).

Four papers have been grouped under 'Statistical methods', an area that has traditionally been strong in humanities computing. Karen Flikeid

writes about 'Techniques of textual and quantitative analysis in a corpus-based sociolinguistic study of Acadian French'. A large corpus of speech from five Acadian areas in Nova Scotia has been collected, transcribed, and analysed by using a variety of statistical techniques, including factor analysis and multidimensional scaling. There is a careful account of the socio-historical context. The result is a fine illustration of how statistical methods can be used in a revealing way, without overshadowing the object of study. G. Lessard and A. Whitfield examine popular Quebec French elements in three Québécois novels. The focus is on the types of elements that are introduced and on the way the three authors integrate them in their texts, to reflect the socio-cultural background of the characters and provide metadiscursive commentary. In an article on 'Frequent words, authorship, and characterization in Jacobean drama' Thomas B. Horton examines and further develops Burrows' method for the analysis of characterization, using frequent words and multidimensional scaling (Burrows 1987). Effects due to authorship may also be revealed by these methods. The final paper in the section on statistical methods is Etienne Brunet's 'What do statistics tell us?' He gives good examples of different statistical techniques applied to literary texts in the FRANTEXT database. At the same time this well-considered paper draws attention to a number of problems in statistical studies of texts, concerned with the definition of units of measurement and the (lack of) interpretation of the results ('results that are not interpreted by the person who drew them up, lists that are not analysed, and concordances that are not exploited', p. 74).

The second and third sections of the book contain papers on 'Text analysis tools' (two papers) and 'Linguistics' (three papers). Hans van Halteren presents a contextual analysis system in 'The Scholar's Work-desk: A STRIDER case study'. Susan Hockey, Jo Friedman, and John Cooper describe the Oxford Text Searching System, which is designed for users with a limited knowledge of computers. Andrea de Leeuw van Weenen writes about 'Automatic lemmatization of classical Armenian texts'. Arne Jönsson and Lars Ahrenberg present a tagging system using 'directed acyclic graphs' and show how it can be used as a tool for the generation of unification-based grammars. In another linguistics paper B. Elan Dresher describes 'YOUPIE: A parameter-based learning model for metrical phonology'.

A group of five papers is assigned to a section on 'Artificial intelligence and computational linguistics'. Nancy Ide and Jean Véronis deal with the understanding of literary narrative, stressing that current AI story understanding systems require fundamental changes if they are to handle

the processing of literary narrative. Christian Koch proposes a neural network simulator as a tool in studying reader-text interaction. Igor A. Mel'cuk and Alain Polguère outline a 'Meaning-text model for English text generation'. Convinced that linguistic and non-linguistic knowledge should be kept apart in text generation, they define some aspects of a separate linguistic module. It is interesting to note that they operate with a semantic representation consisting of a semantic structure representing the propositional meaning of the sentence to be generated, and a communicative structure defined in terms of theme, rheme, and the like. Examples are given of the mapping between semantic and syntactic representations. Another paper, with the somewhat puzzling title 'An unnatural natural language interface' (the table of contents calls it: 'An unnatural language interface'), is concerned with a mapping in the opposite direction. The authors (Nick Cercone et al., Simon Fraser University, British Columbia) present SystemX, which converts English sentences into an unambiguous formal representation to provide a natural language interface to a database. The advantages claimed as compared with other systems include greater portability and greater ability to handle quantification. The focus of the contributions in the book is on aspects of language and texts, with the exception of a paper by Jim Kippen and Bernard Del, who set up a descriptive and generative model of musical structure. Musical patterns are, however, characterized in terms of sentences and grammars, i.e. drawing on the concepts of linguistic analysis.

Lexical databases are the focus of two papers. Frank Tompa and Darrell Raymond consider the elements and relationships of dictionary databases in their 'Database design for a dynamic dictionary'. Nicoletta Calzolari and Antonio Zampolli describe the development at the University of Pisa of a lexicographer's workstation for integrating lexical databases and text corpora. The paper explains how the resources of 'computational linguistics' and 'literary and linguistic computing' can be combined, two research traditions which have been largely separate in the past but now seem to be converging (as shown in a brief historical sketch included in this paper). In another paper, Jacques Dendien, creator of the FRANTEXT database, formulates a mathematical model for storage and retrieval of textual data, and reports that applications of the model led to great gains in storage space and access speed. The last research paper is a report by Patricia Galloway and Clara Sue Kidwell on the development of a database for the 'Choctaw Land Claims Project', which is concerned with defining the relationship over time between the Choctaw Indians in Mississippi and their land. The special challenge

is how to treat a variety of data types within an overall database structure.

In addition to the research papers, there are two general discussion papers introducing and concluding the book, which were originally keynote addresses at the 'Dynamic Text' conference. They provide a broad perspective from the viewpoint of two eminent and experienced humanities scholars, the literary critic Northrop Frye (who unfortunately did not live to see the book in print) and the archaeologist Jean-Claude Gardin. Though recognizing the straightforwad gains achieved by computer methods, they focus on more elusive matters. Have the advances on the information level led to major changes in the substance of knowledge? Have there been qualitative gains as well as quantitative? What should humanities computing attempt to achieve? Frye suggests that the essential assumptions of critical schools could be brought out by computer modelling. In a similar vein, Gardin argues that the computer should be used to model human reasoning processes in the humanities. To speak with the editor, both 'argue persuasively that the humanities needs [sic] to clarify its goals and methodology' (p. vii), and here the computer could play an important role.

Altogether, *Research in humanities computing 1* is well produced, and provides only the occasional typing error or stylistic infelicity (the following one is a wee bit puzzling: 'a wie degree of stylistic variation', p. 225). Together with the yearbook, it gives insight into the range of work in humanities computing and indicates present trends in research and possible future directions. Judging by the two books, humanities computing is alive and thriving. The editors will no doubt have plenty of material when it is time to publish later volumes.

References

Burrows, John F. 1987. *Computation into criticism: A study of Jane Austen's novels and an experiment in method.* Oxford: Clarendon Press.

Johansson, Stig and Anna-Brita Stenström (eds.). 1991. *English computer corpora: Selected papers and research guide.* Berlin: Mouton de Gruyter.

Lancashire, Ian and Willard McCarty. 1988. *The humanities computing yearbook 1988.* Oxford: Clarendon Press.

Nelleke Oostdijk. *Corpus linguistics and the automatic analysis of English.* Amsterdam, Rodopi, 1991. Reviewed by **Josef Schmied**, University of Bayreuth, Germany.

The book under review summarizes ten years of development in corpus linguistics in what the author herself calls the 'Nijmegen style' (in contrast to the 'Lancaster style' illustrated in Garside/Leech/Sampson, eds., *The computational analysis of English.* London: Longman 1987). This approach is characterized (mainly pp. 58–64) by the interaction of intuitions and empirical data, the use of non-probabilistic automatic tools and the automatic conversion of a formal grammar by means of a parser generator. The book can be divided roughly into three parts.

The first section (pp. 1–55) discusses corpus-linguistic principles, previous language processing systems and corpora in general before concentrating on the development of TOSCA and further projects in Nijmegen, which resulted in the Nijmegen corpus and the (automaticized) Nijmegen grammar.

The Nijmegen corpus is exceptional in so far as it contains texts of 20,000 words, having been compiled for a study of linguistic variation. This is still a bold aspiration, because, in contrast to traditional studies of language variation, the corpus-variational perspective aims at describing 'typical features rather than unique ones' (p. 45); but this is still a long-term plan and is not taken up later in the book. The Nijmegen grammar is an extended affix grammar, a context-free grammar which is supplemented with affixes, meta-rules and predicates, all of which are exemplified convincingly.

The second part (pp. 81–147) deals with the empirical description and automatic analysis of language, in which Oostdijk illustrates the formalization of descriptive rules in coordination and gapping. Here she shows that the rewrite rules of Nijmegen grammar can be successfully applied to quite complex language material, but she is also conscious of possible shortcomings and specific problem cases of coordination. These concrete examples serve to show 'how a formalized description of the syntax of a language may be arrived at and what considerations may play a role' (p. 147).

The third part contains an evaluation of the formalized grammar, an informal one and a standardized assessment. Oostdijk demonstrates the advantages of an inductive, corpus-linguistic approach compared to writing a formal grammar that is not constantly tested on a corpus, because the grammar can be adapted continually to the language reality found in the corpus. The examples clearly show that automatic analysis

breaks down with those constructions whose description often remains implicit in the handbooks of English (like coordination without overt coordinator or punctuation mark, on p. 197) and which might not be taken into consideration at all in a deductive approach.

In more general terms, the performance of the parser varied considerably according to genres (with difficult text features such as sentence length): whereas 88% of the fiction text type was basically analysed correctly, the non-fiction text type was analysed *adequately* only in about 56% of all utterances. With the increase in input length the time needed for parsing increased exponentially – and sentences from the non-fiction sample did not result in any successful analyses after 1800 CPU-seconds (p. 216). Thus Oostdijk also shows that – although this may give rise to problems of consistency – a human analyst must perform some necessary tasks: in a syntactic pre-analysis certain problematic constituents have to be marked, such as noun phrases in non-typical functions (which are listed in appendix H).

Although the three parts are somewhat heterogeneous and the subject matter does at times become rather complicated, the volume as a whole reads very well. This can be attributed to its relatively fluent style. The description is openly realistic, and despite her personal involvement in its development the author tries to assess the project objectively in informal as well as formal terms. The language and argumentation is relatively easy to follow from a linguistic perspective, obviously because Oostdijk sees computational technology 'merely as a means to an end' (p. 2) and tries to 'stay/keep close to what is traditional and familiar' (p. 205 and 149) consistently in her grammatical terminology.

Unfortunately, she uses *A grammar of contemporary English* (1972) as a basis for the description of coordination and gapping (ch. 4) and does not take into account some useful expansions in *A comprehensive grammar of the English language* (1985), because this part of her work had basically already been written before the publication of the latter. Similarly, Biber's book (*Variation across speech and writing*. Cambridge: C.U.P. 1988) could have been used for the description of his multi-feature/multi-dimensional approach (pp. 40-44) instead of the previous articles alone.

From a corpus-linguistic perspective some linguists might suggest that the principles of the subdiscipline would have been adhered to even more consistently, if the sample sentences had been taken from real utterances in the corpus analysed. But despite these minor points the volume succeeds in showing that corpus linguistics 'has developed into

a discipline in its own right, while continuing to be an important ancillary discipline to various other linguistic subdisciplines' (preface).

Felicitas Tesch. *Die Indefinitpronomina some und any im autentischen englischen Sprachgebrauch und in Lehrwerken.* Tübinger Beiträge zu Linguistik 345. Tübingen: Gunter Narr Verlag. 1990. Reviewed by **Magnus Ljung,** University of Stockholm.

The aim of this study is twofold. The book's more immediate aim is to compare the use of *some* and *any* and their combinations in (certain parts of) the LOB and London-Lund corpora with that in certain German textbooks. The ultimate aim, however, is to construct a partial didactic grammar for *some* and *any* on the basis of these data.

In the analysis, two subsections from the LOB Corpus are used, i.e. the non-fictional LOB_S which comprises text categories A, B, C, F and G (all in all 184,000 words), and the fictional LOB_F containing 176,000 words from categories K; L, M and P. From the London-Lund Corpus – CONV – the 34 surreptitiously recorded texts have been used, i.e. 170,000 words.

The textbook category contains six books from *The Learning English Modern Course: Gym.* We are told that it consists of six volumes, but so far as I can see, no mention is made of the number of words in the textbook corpus.

The book has three parts, one theoretical and two empirical. The theoretical part contains a survey of previous literature, a presentation of the aims of the study, and a description of the corpus and its structure. We are also given a detailed description of the preparatory work with the corpus, the variables used and even the coding plan used in the programmes.

The second part contains the results from the analysis of the LOB and London-Lund corpora and the third a comparison between the authentic data and that found in the textbooks.

Tesch's findings concerning the use of *some* in the authentic texts agree with traditional accounts of *some*. With regard to *any*, she distinguishes between three types on the basis of their semantic, syntactic and pragmatic properties. *Any*-1 has the meaning 'all', 'any... what-(who-, which-)ever' (*jeder beliebige* p.343) and presupposes the existence of its referent. It occurs in affirmative declaratives and in interrogatives without negation. This type of *any* makes up 50% of all *any*-instances in Tesch's authentic data.

Any-2 occurs in declarative negative sentences and makes up between 30% and 40% of all cases of *any*. The third type *Any*-3 is found in affirmative interrogative contexts. Neither of the last two types carry presuppositions about the existence of the referents.

In the comparison between the treatment of *some*- and *any*-forms in the authentic and the didactic texts, Tesch finds no discrepancies for *some*. However, in the authentic texts the distribution of *any*-1 across declarative and interrogative contexts was 80:20, while in the textbooks the ratio turned out to be 50:50. In these texts, accordingly, the use of *any*-1 in interrogatives is thus clearly over-represented. The book ends with two suggestions for how the rules for *some* and *any* should be formulated, in which the prominence of *any*-1 is given full recognition.

These results are interesting, since in traditional accounts *any* is primarily associated with negative contexts. Tesch's results are also borne out by other data: a small study of the *any* forms in 50,000 words of *Newsweek* texts from 1989 yielded the same distribution over affirmative and negative contexts.

Several other interesting facts are unearthed in Tesch's study. She finds, for instance, that *some*- and *any*-words – i e both the simple and compound forms – are twice as common in the spoken as in the written language, but the ratio between *some*- and *any*-words is roughly equivalent in both the spoken language and in the two written corpora (to be exact, about 55% *some* vs 45% *any* in CONV and about 60% *some* and 40% *any* in LOB_s and LOB_f).

She also finds that in both CONV and LOB_f the rank list for the different functions of these forms is (1) pronoun (2) determiner (3) adverb, while for LOB_s it is (1) determiner (2) pronoun (3) adverb.

Another finding is that in all three corpora the simple forms *some* and *any* are more frequent than the combined forms and that the compounds in *-one* are much more common than the ones in *-body* in the written language, while the reverse is true of the spoken texts.

A great deal of work has obviously gone into this investigation and in many respects it is a good example of a nofrills down-to-earth corpus study. The organization of the book is reasonably straightforward, although it is sometimes rather difficult to find authentic examples, for example of the three *any*-types.

On the other hand there is at times an excess of documentation: it is hardly necessary in a book of this type to explain the workings of the chi-square calculation in detail or to give a detailed account of the coding principles.

It is unfortunate, in my opinion, that Tesch decides against including any discussion of phonology. The fact that most of the data comes from written sources does not mean that matters like stress and strong/weak forms cannot be brought to bear on the definitions of the different types of *some* and *any*. In particular, it strikes me as odd to formulate a rule for teaching purposes which contains no mention of either stress or pronunciation.

All in all, however, Tesch has made a valuable contribution to the field of corpus studies, which offers a wealth of information about the use of *some* and *any* in authentic texts.

Gunnel Tottie. *Negation in English speech and writing: A study in variation.* San Diego & London: Academic Press, 1991. ISBN 0-12-696130-1. 353 pp. Reviewed by **Leiv Egil Breivik**, University of Bergen.

Over the centuries, the study of negation has enjoyed considerable popularity among philosophers and linguists. Problems which have been amply discussed in the literature include negative scope *(All the boys didn't leave)*, *neg*-raising *(I don't think he's coming)*, double or multiple negation *(I don't know nothing no more)*, and the use of negative polarity items (cf. the variation between *some* and *any*). One of the most influential twentieth-century contributions to the field is Jespersen's (1917) monograph on the use of negation in English and other languages. Among more recent works, particular mention should be made of Klima's (1964) attempt to present a unified treatment of the entire system of negation in English within the framework of early transformational-generative syntax. Klima's seminal article spawned a great number of studies dealing with different aspects of negation, highlighting some of the main issues of generative grammar.

As is well known to readers of this journal, Tottie has been a prolific writer on the subject of negation in English (e.g. Tottie 1977, 1982, 1985, 1987). The book under review, which is a synthesis and extension of her previous work, represents the most comprehensive and systematic investigation to date of negation in English; it focuses on problems which have not been given much attention by other scholars. The author's primary concern is to set up a pragmatic theory of the use of negation in English and to examine in detail two types of morphosyntactic variation: first, the variation between affixal and non-affixal forms (as in *It is impossible / It is not possible*), and second, the variation between *not*-negation and *no*-negation (as in *He did not see anything / He saw nothing*).

Negation in English Speech and Writing is organized into eleven chapters. The introductory chapter (pp. 1-14) contains a description of aims, scope, methodology and material, as well as a brief survey of earlier work on negation. The overall approach adopted by Tottie is empirical and quantitative: 'the theoretical framework is that of the quantitative study of linguistic variation and variation theory as laid down principally in the works of William Labov. I thus base my work on the large-scale study of attested empirical data, and I analyze the use of variant forms to convey the same meaning, with the purpose of determining the factors which condition their use' (p. 10). In other words, Tottie posits that there should be semantic equivalence between alternating forms. All of the spoken material is taken from the London-Lund Corpus of Spoken English and is made up of casual conversation. The written material, consisting of expository prose, derives from two sources: the files of the Survey of English Usage at University College London and the Lancaster-Oslo/Bergen (LOB) Corpus of British English. The size of the corpus varies somewhat between the different chapters.

In Chapter 2 (pp. 15-30), the first step is taken towards developing a theory which accounts for the different uses of negation in spoken and written English. Tottie's point of departure is that in her material, the average frequency of negative expressions is more than twice as high in the spoken texts as in the written texts (27.6 vs. 12.8 per 1,000 words). She proposes the following discourse-functional classification of negative sentences:

(i) rejections (including refusals)
(ii) denials: (a) explicit
(b) implicit

Denial relates to propositions and normally requires linguistic means for expression, while rejection is essentially a pragmatic category and thus not dependent on language (although it may be expressed in natural language). Explicit denials refer to denials of explicit assertions; implicit denials deny propositions that can be inferred from the co(n)text but which have not been explicitly asserted. Consider the following conversational exchanges:

(1) (a) X: Would you care for some wine?
(b) Y: No thanks, I don't drink.
(2) (a) X: Come and play ball with me.
(b) Y: No, I don't want to.
(3) (a) X: John is married.
(b) Y: John isn't (married).

(4) (a) X: John's wife is a teacher.
(b) Y: John isn't even married.

In (1b) speaker Y rejects what X has offered him. In (2b) Y refuses to do what X has just suggested. In (3b) Y denies what has been explicitly stated by X, while in (4b) Y merely denies the presupposition of (4a).

Tottie hypothesizes that the above framework will account for the much higher incidence of negative expressions in speech than in writing. Thus she finds it plausible that rejections (of offers etc.) and explicit denials of propositions are more natural in conversation, where the interlocutors are physically present. In Chapter 3 (pp. 31-44), the framework is tested against quantitative findings, and it is shown that although rejections and explicit denials are indeed more frequent in speech than in writing, there are a number of other factors which contribute substantially to the high frequency of negation in conversation. The most important of these factors is the tendency for negative forms to collocate with mental verbs like *know* and *think*, which 'are typically used with negation as face-saving devices, i.e. to render, for example, a rejection more palatable to the hearer, as in *I don't know # I don't know whether I'll drink coffee at this time of day*' (p. 315). Furthermore, negative expressions occur in repetition and questions, and they are used as support signals in conversation.

The bulk of the book deals with variation along the two dimensions mentioned in the introductory paragraphs. Chapters 4-5 (pp. 45-85) examine the variation between affixal and non-affixal variation, while Chapters 6-10 (pp. 87-312) are devoted to the variation between *not*-negation and *no*-negation. The author here combines a variationist approach with a discourse – analytic perspective in an attempt to arrive at the factors which determine the use of alternating forms. Historical data are also cited to shed light on the present-day situation.

The most important result which emerges from the investigation reported in Chapters 4-5 is that affixal negation of adjectives is much commoner in the written than in the spoken material; in the former sample about two-thirds of the total number of negative sentences with adjectives have affixal negation, whereas in the latter only one-third exhibit this variant. Tottie attributes this to the use of different discourse strategies in speech and writing: 'Because of the greater pressure imposed on speakers, they tend to produce utterances where one idea follows another in a **fragmented** discourse, whereas writers typically have more time to combine and superimpose ideas on each other and can therefore mold

their thoughts into a more **integrated** discourse (cf. Chafe, 1982)' (p. 317). It should be pointed out, finally, that variation between affixal and non-affixal variation is constrained by several factors ('knockout constraints'), e.g. by lexical gaps (as in *intact / *not tact)* and by the presence of adverbials (as in *it was absolutely illogical ≠ it was not absolutely logical).* (It would have enhanced the value of Tottie's discussion of the semantic differences between affixal and non-affixal negation (pp. 50-55) if she had taken account of Rusiecki's (1985) important treatment of gradable adjectives.)

The distribution of *not*-negation and *no*-negation is also significantly different in the two samples. In speech, there is 66% *not*-negation and 34% *no*-negation. The proportion is almost exactly reversed in the written material: 37% *not*-negation and 63% *no*-negation. Several individual factors influence the choice of *not*-negation or *no*-negation. These factors are of two main types, **global** and **specific.** Global factors consist of, or affect, more than one sentence element, whereas specific factors consist of either the verb phrase or the *neg*-incorporating element. Tottie uses a variable rule program, VARBRUL 2S, to establish the influence of each factor on the choice of negation type. Her analysis reveals, for example, that existential *be* favours *no*-negation in both speech (.838 probability) and writing (.907), whereas the copula *be* disfavours this variant in both samples (.079 probability in speech and .112 in writing). Other factors which are tested by the variable rule analysis include verb-phrase complexity (simple or complex), type of incorporating element (e.g. adverb, adjective or pronoun) and type of pronoun (e.g. *-thing* or *-body).*

There can be no doubt that the variation between *not*-negation and *no*-negation in present-day English is conditioned by a conglomerate of global and specific factors. Tottie argues, to my mind convincingly, that the effects of the individual factors can only be understood if seen against the backdrop of the diachronic development of negation in English. The overall difference between speech and writing is not surprising, she contends, in view of the fact that *no*-negation antedated *not*-negation. (While the precursor of *no* (i.e. *ne*) already occurred in Old English, *not* did not become common as a sentence adverb until the Middle English period.) She writes (p. 325): 'the likeliest explanation for the discrepancy between spoken and written English is that the spoken variety represents a more advanced stage of the language and writing a more conservative type. The present situation is no doubt a stage in a long-term development, where the locus of change is as usual speech, and where innovations slowly trickle into the written variety'.

Frequency is claimed to play a major role in this development, e.g. in the preservation of *no*-negation in sentences with high-frequency lexical items like *have*, *never* and *nothing*; less frequent items such as most lexical verbs and most nouns prefer *not*-negation, 'presumably because they were not as frequently used in collocations with *no*-negation' (p. 326). Interestingly, if Tottie's explanation is correct, the development of negation in English provides evidence for lexical diffusion in syntax. However, as the author points out herself, much more work remains to be done before we can establish the validity of the diachronic-synchronic hypothesis she proposes.

The final chapter (pp. 313-31) summarizes the findings presented in the preceding chapters, considers how they can be accounted for and discusses some of their implications for linguistic theory. This chapter contains a number of interesting claims and observations (some of which have been anticipated above). However, apart from Tottie's wishing to set up a framework which refers to both diachronic and synchronic parameters (a goal with which I am broadly in sympathy; cf. Breivik 1989), she does not discuss the full implications for grammatical description of the position she adopts (indeed, the full implications are not strictly relevant to her specific purpose). In the absence of any such discussion, one could well imagine the theoretical background to the discussion of the diachrony of negation to be something like Givón's functional-typological approach (see e.g. Givón 1984, 1990).

The book under review represents an invaluable source of data and is rich in insightful discussions and analyses. The evidence Tottie provides for her hypotheses cannot be easily dismissed. The presentation is lucid throughout. The extensive subject and author indexes at the end also enable the reader to use the book effectively as a reference work. Any scholar – of whatever theoretical persuasion – interested in negation in English must take note of *Negation in English Speech and Writing*. Certain innovations of technique, as well as the overall formulation of Tottie's attack on the problem of linguistic variation, make this work merit the attention even of those whose interests lie outside the area of negation in English.

References

Breivik, Leiv Egil. 1989. On the causes of syntactic change in English. In *Language change: Contributions to the study of its causes*, ed. by Leiv Egil Breivik and Ernst Håkon Jahr. 29-70. Berlin & New York: Mouton de Gruyter.

Chafe, Wallace. 1982. Integration and involvement in speaking, writing, and oral literature. In *Spoken and written language*, ed. by Deborah Tannen. 35-53. Norwood, N.J.: Ablex.

Givón, Talmy. 1984-1990. *Syntax: A functional-typological introduction.* 2 vols. Amsterdam & Philadelphia: Benjamins.

Jespersen, Otto. 1917. *Negation in English and other languages.* Copenhagen: Det Kgl. Danske Videnskabernes Selskab.

Klima, Edward S. Negation in English. In *The structure of language*, ed. by Jerry A. Fodor & Jerrold J. Katz. 246-323. Englewood Cliffs: Prentice-Hall.

Rusiecki, Jan 1985. *Adjectives and comparison in English: A semantic study.* London: Longman.

Tottie, Gunnel. 1977. *Fuzzy negation in English and Swedish.* Stockholm Studies in English, 39. Stockholm: Almqvist & Wiksell.

Tottie, Gunnel. 1982. Where do negative sentences come from? *Studia Linguistica* 36:88-105.

Tottie, Gunnel. 1985 The negation of epistemic necessity in British and American English. *English World-Wide* 6:87-116.

Tottie, Gunnel. 1987. Rejections, denials, and explanatory statements: A reply to Fretheim. *Studia Linguistica* 41:154-63.

Conference reports

Twelfth ICAME Conference, 1–12 May 1991

Lou Burnard
Oxford University Computing Service

The 1991 ICAME conference was hosted by Leeds University at a splendid Victorian hotel on the edge of Ilkley Moor and enjoyed excellent weather, the usual relaxed atmosphere and the usual extraordinary array of research reports, which can only be very briefly noticed in this report. As usual, there were about 50 invited delegates, most of whom knew each other well. The social programme included an outing to historic Haworth by steam train.

For the first time, the organising committee had included a so-called open day, to which a number of interested parties had been invited. As curtain raiser to this event, I was invited to present a status report on the Text Encoding Initiative (TEI) and Jeremy Clear (OUP) to describe the British National Corpus project. The open day itself included brief presentations from Stig Johansson (Oslo), on the history of ICAME since its foundation in 1977, from Antoinette Renouf (Birmingham) on basic design problems in corpus building, from Sid Greenbaum (London) on the design and implementation of the new co-operative International Corpus of English project, from Eric Atwell (Leeds) on the kinds of parsing systems which corpus linguistics makes possible, from Jan Aarts (Nijmegen) on the Nijmegen approach to computational linguistics, from John Sinclair (Birmingham) on the revolutionary effect of corpus linguistics on lexicography and on language teaching, from Gerry Knowles (Lancaster) on the particular problems of representing spoken language in a corpus and from Knut Hofland (Bergen) on the technical services provided for ICAME at Bergen.

The conference proper began with a series of papers about electronic

lexica of various flavours, ranging from the CELEX database (Nijmegen) in which a vast array of information about three languages (Dutch, English and German) is stored in a relational database, to the experimental word-sense lattices traced by Willem Meijs' Amsterdam research team from the LDOCE definitions. Work based on this dictionary was also described by Jacques Noël (Liège) and by Louise Guthrie (NMSU). The former had been comparing word-senses in Cobuild and LDOCE, while the latter had been trying to distinguish word senses by collocative evidence from the LDOCE definition texts.

The traditional ICAME researcher first quantifies some unsuspected pattern of variation in linguistic usage and then speculates as to its causes. Karin Aijmer (Lund), for example, reported on various kinds of 'openers' in the 100 or so telephone conversations in the London-Lund Corpus, in an attempt to identify 'routinised' patterns. Bengt Altenberg (Lund) reported on a frequency analysis of recurrent word class combinations in the same corpus, and Pieter de Haan (Nijmegen) on patterns of sentence length occurrences within various kinds of written texts.

Two immaculately designed and presented papers concerned work at the boundary between speech as recorded by an acoustic trace and by transcription: Anne Wichmann (IBM) presented an analysis of 'falls' in the London-Lund Corpus, and Gerry Knowles (Lancaster) proposed a model for speech transcription.

High spots of the conference were the presentations from Tim O'Donoghue (Leeds) and Mitch Marcus (University of Pennsylvania). If there is anyone around who still does not believe in systemic functional grammar, Tim O'Donoghue's presentation should have converted him or her. He reported the results of comparing statistical properties of a set of parse-trees randomly generated from the systemic grammar developed by Fawcett and Tucker for the Polytechnic of Wales Corpus with the parse trees found in the same (hand-)parsed corpus. The high degree of semantic knowledge in the grammar was cited to explain some very close correlations while some equally large disparities were attributed to the specialised nature of the texts in the corpus. Mitch Marcus gave a whirlwind tour of the new burgeoning of corpus linguistics in the US. He described the methods and design goals of the Penn Treebank project, stressing its engineering aspects and providing some very impressive statistics about its performance.

Several presentations and one evening discussion session concerned the new 'International Corpus of English' or ICE project. Laurie Bauer (Victoria University) described its New Zealand component in one presentation, while Chuck Meyer (UMass) described some software

developed to tag it (using Interleaf) in another. The most interesting of these was from And Rosta (London) who is largely responsible for ICE's original encoding scheme.

There was a general feeling that standardisation of linguistic annotation was long overdue. Marcus pointed out that the Brown Corpus had used 87 different tags for part of speech, LOB 135, the new UCREL set 166 and the London Lund Corpus 197. In Nijmegen, the TOSCA group has an entirely different tagset of around 200 items which has been adopted and, inevitably, increased by the ICE project. It would be a good idea, it seems, for someone to try to see whether these various tag sets can in fact be harmonised using the TEI recommendations.

The 'Using Corpora' Conference, Oxford 1991

Stig Johansson
University of Oslo

The beginning of the 1990s has seen the launching of corpus collection initiatives on a scale which far surpasses that of most previous corpus projects. It was therefore appropriate that 'using corpora' was chosen as the theme of the Seventh Annual Conference of the UW Centre for the New OED and Text Research, held at St. Catherine's College, Oxford, on 29 September – 1 October, 1991. Like its predecessors, this conference was attended by a large number of participants from universities, publishers, and industry. It maintained the high standard we have been used to from previous conferences in the series (to the credit of Timothy Benbow, the Conference Chair, and Frank Tompa, the Nominating Committee Chair).

The programme included theoretical papers, research reports, and general discussions of corpus matters. Most of the papers had a connection with the study of lexis, as is natural at a conference connected with the OED. Papers by Marti Hearst and Adam Kilgarriff dealt with word sense disambiguation. John Justeson presented a paper (written in collaboration with Slava Katz) on the relation of antonomy based on the study of antonymous adjectives in large corpora. Sam Coates-Stephens examined an approach to the analysis of proper nouns in news texts.

A research group from the University of Pisa (Remo Bindi et al.) were concerned with the development of statistical techniques for extracting lexical information from corpora, to be integrated with information from machine-readable dictionaries and linguists'/lexicographers' knowledge within a computational lexicon. Julia Pajzs reported on the use of a lemmatized corpus in compiling a dictionary of Hungarian. Frank Knowles presented some statistical methods for characterizing vocabulary, with reference to the Ashton Corpus of Soviet Yiddish. Stephen Bullon and Tim Lane described their work on a World Service Dictionary, based on data from the BBC World Service.

That corpora have a range of uses, not only in lexical studies, was well shown in papers by Pieter de Haan and Bengt Altenberg on corpus projects at the universities of Nijmegen and Lund. The former focused on the syntactic analysis of corpus data, the latter on the study of spoken texts, especially recurrent lexical patterns and the relationship between grammar and intonation, with applications in computational linguistics, speech technology, lexicography, and language teaching.

Two papers dealt with bilingual corpora, which may become powerful tools in bilingual lexicography, contrastive analysis, and translation studies (including the development of machine translation systems). Elisabetta Marinai, Carol Peters, and Eugenio Picchi described a system for the automatic creation and retrieval of parallel concordances from bilingual corpora. Kenneth Church (speaking on behalf of himself and William Gale) described different tools for establishing word correspondences in parallel texts. In another contribution Kenneth Church (on behalf of himself and Mark Liberman) gave a status report on the Data Collection Initiative of the Association for Computational Linguistics, with special reference to the release of a CD-ROM from the project (see the note in this journal, p. 139).

Most of the papers presented at the conference are reproduced in the Proceedings, which can be ordered from the UW Centre for the New OED and Text Research, University of Waterloo, or from the Dictionary Department, Oxford University Press. As the Proceedings were published before the conference, they could not include reports on two of the most stimulating sessions during the conference, a discussion of corpus matters by John Sinclair and Jeremy Clear, and the concluding Oxford debate on the motion 'A corpus should consist of a balanced and representative selection of texts'. The latter in particular requires some comment. After a lively debate, with Sir Randolph Quirk and Geoffrey Leech speaking for the motion and John Sinclair and Willem Meijs against, the motion was defeated. The result is understandable, in view

of the development in the last three decades from fairly small, carefully constructed corpora to the vast data collections which are now becoming available. Nevertheless, there is still something to be said for the small, carefully constructed corpus which samples from a variety of text types (not just those which happen to be easily available) and can be subjected to total accountability, forcing researchers to see what they might otherwise overlook. The vast data collections, unless they are used systematically to retrieve a subcorpus that is relevant for a particular research project, invite broad quantitative investigations rather than delicate, qualitative studies. We need both types of corpus initiatives, and proponents of both approaches must admit that a corpus, however large or however well balanced it is, may not by itself provide sufficient evidence relevant to a particular research question. The uses of corpora are indeed wide and varied, but we must not forget the limitations.

Nobel Symposium on Corpus Linguistics, Stockholm, 4–8 August 1991

Göran Kjellmer
University of Göteborg

'Nobel Symposia' are arranged by the Swedish Nobel Foundation as a recurrent feature of the intellectual life of Sweden. Over the years, they have dealt with a great number of subjects, all of them considered to be of pioneering importance in fields of scholarly, scientific or technological research. In August 1991, a Nobel Symposium was devoted to Corpus Linguistics. The proceedings took place, and the participants were lodged, at the IBM Nordic Education Center on the island of Lidingö just outside Stockholm. Practical matters were seen to and problems solved by the organiser-in-chief, Jan Svartvik of Lund University, in cooperation with an organising committee and representatives of the Nobel Foundation and IBM.

A number of prominent linguists, both professed corpus linguists and others, had been invited to submit papers on some aspect of corpus linguistics well in advance of the Symposium, and another group of linguists had been invited to prerare comments on those papers. At the Symposium, the presentation of each paper (by a 'Speaker') was followed

first by the comment from a 'Commentator' and then by a free discussion. This proved to be a productive arrangement; the sessions were generally very stimulating with often quite lively exhanges. In accordance with the recommendations of the Nobel Foundation, there was also a session that was open to the public; at this session, which took place at the Royal Swedish Academy of Sciences in Stockholm, Charles Fillmore and M.A.K. Halliday each gave a lecture.

As could be expected with the above arrangement, the programme provided very substantial food for thought. The emphasis was naturally on theoretical issues. Broadly speaking, the topics fell in the following areas:

History of corpus linguistics

Speaker: W. Nelson Francis

Relation of corpus linguistics to language and linguistic theory

Speakers:
Wallace Chafe
Charles J. Fillmore
M.A.K. Halliday
Geoffrey Leech

Principles of corpus analysis

Speakers:
Jane A. Edwards
Geoffrey Sampson
John M. Sinclair

Survey of corpora

Speaker: Martin Gellerstam

Projected corpora

Speakers:
Sidney Greenbaum
Sir Randolph Quirk

Use of corpora in specific domains

Speakers:
Douglas Biber (domain: referential strategies)
Ruqaiya Hasan (domain: rationality in interchange between mothers and children)

Staffan Hellberg (domain: the Swedish Academy Grammar)
Graeme Kennedy (domain: language teaching)
Henry Kučera (domain: production of language aids)
Matti Rissanen (domain: historical studies)

In addition there were demonstrations of computer software by Benny Brodda and Fred Karlsson. The Proceedings of the Symposium will be published in *Directions in corpus linguistics* (ed. by Jan Svartvik, Berlin: Mouton de Gruyter).

Social events included a performance by local folk dancers and visits to the Vasa museum, which houses a 17th-century warship, and to Millesgården, a permanent exhibition of the work of the 20th-century Swedish artist Carl Milles.

The papers and discussions at the Symposium conveyed a strong impression not only that corpus linguistics has finally established itself in the scholarly arena but also that there are untold exciting research tasks just round the corner. According to the directives of the Nobel Foundation, Nobel symposia are to be devoted to 'break-through research areas and to topics that are considered to be great future relevance to society'. The participants in the August of 1991 Symposium must have felt that this is an adequate description of corpus linguistics today.

Seminar on Corpus Studies and the Computer in English Language Research, Tampere 1991

Anna-Brita Stenström
University of Bergen

Ian Gurney, lecturer in the department of English at Tampere University, Finland, organized their third national research seminar 21-22 November 1991. The seminar was devoted to the use of computers in corpus-based studies of English and aimed at students and researchers from Tampere and other Finnish universities. Speakers had been invited from Britain, Sweden, Norway and Finland.

The papers gave a good insight into the different types of corpus research that can be carried out with computer assistance, both diachronic

(eg Merja Kytö's paper on 'The Helsinki Corpus as a tool in diachronic research') and synchronic (eg Ian Gurney's paper on 'Zero genitive or *s*-genitive: pluralities and singularities in usage').

The well organized and very fruitful seminar, which also included software demonstrations, ended with a discussion where certain problems connected with this type of research were considered. How, for instance, does one instruct the computer to retrieve all the cleft sentences in a Middle English text?

Shorter notices

LOB – 30 years on ...

Andrea Sand and Rainer Siemund
Freiburg University

Work is under way at Freiburg University (FRG) to compile a corpus of written British English to match the Lancaster-Oslo/Bergen Corpus as closely as possible in size and composition. The aim of the project is to statistically document linguistic innovation and changing stylistic norms in the present-day language.

On hearing of such a project many of those concerned with corpus linguistics might ask why take trouble to compile a one million word collection at a time when corpora a hundred times as large will be widely available in the forseeable future. In short, we see two advantages of corpora compiled along the lines of Brown and LOB which seem to justify our effort. The first is their wide coverage of many textual genres and stylistic registers of written English (– we dare not use the term 'representativeness'). The second is their availability for teaching – an advantage that is being used extensively in Freiburg already.

Like its 1961 counterpart the new corpus will contain 500 texts of about 2000 words each, roughly a million words in all. Over the last few months we have compiled a first part matching the 'press' component to test the feasibility of the project. 'Press' contains about 176,000 words, incorporating the categories Reportage, Editorial and Reviews. These again are subdivided into national and regional papers, daily and weekly publications.

To keep as close as possible to LOB, we made an effort to use the same papers that had been chosen by the compilers of the original. This proved to be difficult in many cases though. The media landscape, needless to say, has changed quite a lot over the last thirty years. As we learned from *Willing's Press Guide*, our main source of information, quite a few papers had ceased publication in the meantime. So we had

to select a number of new papers which we grouped around the core of 'veterans'. We deliberately excluded papers which are circulated in vast quantities without charge. Although they are a 'sign of the times' we ranked the comparability of LOB '91 to LOB '61 higher in priority than the possible alternative goal, viz. to create the accurate picture of the British printed press right now. In some cases, however, papers simply were not willing to co-operate so we had to exclude them. For all these reasons we ended up with a slightly different distribution from the original LOB, as illustrated on maps 1 and 2. Once we had chosen our papers, new problems arose. In some cases, publications had been continued, but do not feature certain text categories any more. Many papers, for example, still include institutional editorials, while the personal editorial is in decline. The latter was the most problematic text type and we decided to replace it partly by institutional editorial, which is similar in language but used more frequently. Generally, we tried to shift texts only within text categories.

Other difficulties arose with the categories 'political reportage', 'cultural reportage' and 'reviews'. The political section of very many papers has been reduced to a 'news-in-brief' part, which mainly consists of the pre-fab material from international news agencies. The situation is even more problematic in the cultural sections. Most of the reviews we found in other than quality national papers – in case there were any at all – are mainly concerned with recently released videos or the latest hit album. For the sake of conformity in the newspaper selection we stretched the concepts of 'political reportage' and 'culture' as far as possible. This also throws light on what is regarded as culture by most publishers and readers.

Our original plan to scan the newspaper material into our computers failed, mainly because of the columns, smudges and ligatures which are characteristic for newspaper texts. The editing of this kind of scanned data took much longer than the typing of the complete original. Here we hope for better results with the other sections of the corpus.

For further information contact:

Prof. Dr. Christian Mair
Englisches Seminar I
Institut für Englische Sprache und Literatur
Albert-Ludwigs-Universität
Postfach
D-7800 Freiburg i. Br.

We invite interested researchers to correspond with us through the above address and on completion of the project will distribute the 'new LOB' eventually through ICAME.

Map 1: LOB 1961

Map 2: LOB 1991

Building a million-word computer science corpus in Hong Kong

Robert Davison
Hong Kong University

The aim of this project is to compile a corpus of one million words in the field of Computer Science, books for sampling having been chosen from First Year university reading lists, primarily at the Hong Kong University of Science and Technology.

Texts are taken in 2,000-word chunks, selected at random from the first, second and third 'thirds' of each book and then photocopied, scanned and transferred to WordPerfect for spell-checking and proof-reading. The definition of computer science has proved to be problematic, and without detailed analysis here, it has been taken to cover such fields as: networking, databases, expert systems, programming languages, artificial intelligence, etc.

The research was established on the rationale that students learning in a second language, as in the case in Hong Kong, where their mother tongue is Chinese (Cantonese) and yet the language of tertiary education is often English, often experience great difficulties with the language itself, especially non-technical language. When students encounter an unknown word, they will turn to a dictionary and tend to choose the very first explanation that the dictionary offers. Yet all too often this is the incorrect meaning and hence they misunderstand and meaning of the text. Thus, having collected a sufficient body of text, analysis of usage can be carried out on the *sub-technical* features of the language: not the computer science terms such as 'database', 'network' and 'system', but non-technical words which are often used in such technical texts. Eventually, an annotated glossary of how to use computer science words in English can be produced to help students with comprehending such texts. However, feedback we have recently received from seminar and conference participants has suggested another way of using the corpus, namely, compiling a context-specific grammar for the subject of computer science. In this respect we are still very much open to ideas about how the material can be most effectively used, though evidently the primary focus is on the development of materials which can be used, in one way or another, for teaching and learning.

For more information contact:

Robert Davison
Research Assistent
Language Centre
Hong Kong University of
Science and Technology
Clear Water Bay
Hong Kong
E-mail: lcrobert@usthk.bitnet

Dr. G. James
Director
Language Centre
Hong Kong University of
Science and Technology
Clear Water Bay
Hong Kong
E-mail:lcgjames@usthk.bitnet

The Lancaster Parsed Corpus

Geoffrey Leech
Lancaster University

This is a parsed subcorpus of the Lancaster-Oslo/Bergen (LOB) Corpus, compiled by Roger Garside, Geoffrey Leech and Tamás Váradi. It can now be obtained (under conditions similar to those applying to other corpus holdings) through ICAME.

The Lancaster Parsed Corpus is a treebank consisting of sentences of the LOB Corpus, amounting altogether to over 133,000 words. Each sentence in the Parsed Corpus is annotated with a phrase-structure parse, represented in the form of labelled bracketing, marking the boundaries of sentence, clause, phrase, and coordinated word constituents. The labels correspond to well-known 'consensual' constituents such as noun phrases, relative clauses, infinitive clauses, etc. The annotations also include the word tags used for the Tagged LOB Corpus. See the examples on p. 125.

The corpus consists of syntactically analysed sentences from each text category of the LOB Corpus, viz. A: Press, reportage; B: Press, editorial; C: Press, reviews; D: Religion; E: Skills, trades, and hobbies; F: popular lore; G: Belles letters, biography, essays; H: Miscellaneous; J: Learned and scientific writings; K: General fiction; L: Mystery and detective fiction; M: Science fiction; N: Adventure and western fiction; P: Romance and love story; R: Humour.

The average length of sentences in the Lancaster Parsed Corpus is c. 11 words, considerably less than the average length of sentences for

Sample sentences from the Lancaster Parsed Corpus

A07 418

```
[S[Na I_PPlA Na] [V can_MD n't_XNOT make_VB V][N a_AT club_NN N][Tb[V
pay_VB V] [N a_AT player_NN N][N[D so_QL much_AP D][N a_AT week_NN N]N]
Tb]._. S]
```

B04 248

```
[S[N \OHr_NPT Henry_NP Newton_NP [Po of_INO [N Acton_NP N] Po]N][V does_DOZ
not_XNOT want_VB V][N his_PP$ daughter_NN N] [Ti [Vi to_TO marry_VB Vi][N
a_AT Scotsman_NNP N]Ti] ._. S]
```

D03 191

```
[S[N the_ATI word_NN [P for_IN [N guiding_NC N]P]N][V comes_VBZ V][P
from_IN [N a_AT root_NN [Fr[Nq that_WP Nq][V is_BEZ used_VBN V][Ti[Vi to_TO
describe_VB Vi][N& the_ATI herding_NN [Po of_INO [N sheep_NNS N]Po][N+
or_CC the_ATI conducting_NN [Po of_INO [N ' prisoners_NNS
N]Po]N+]N&]Ti]Fr]N]P] ._. S]
```

E01 2

```
[S[Nq what_WDT Nq][N a_AT world_NN [Po of_INO [N graceful_JJ
accomplishment_NN N]Po]N][V lies_VBZ V][P in_IN [N a_AT piece_NN [Po of_INO
[N[J finely_RB worked_VBN J] hand-made_JJ lace_NN N]Po]N]P] [_! S]
```

G09 264

```
[S[N this_DT N][V was_BEDZ V][N all_ABN part_NN [Po of_INO [N the_ATI
act_NN N]Po]N] ._. S]
```

H04 158

```
[S[N table_NN 26_CD N][V gives_VBZ V][N similar_JJ regression_NN
estimates_NNS [P for_IN [N previous_JJ years_NNS N]P]N] ._. S]
```

J03 113

```
[S& [N their_PP$ orientation_NN N] [V will_MD give_VB V][N a_AT sense_NN
[Po of_INO [N the_ATI direction_NN [Po of_INO [N movement_NN
N]Po]N]Po]N][S+ and_CC [R often_RB R][N a_AT good_JJ deal_NN N][V can_MD
be_BE learned_VBN V][P from_IN [N the_ATI kind_NN [Po of_INO [N stone_NN
N]Po]N]P]S+] ._. S&]
```

L03 284

```
[S[N Grundy_NP N][V would_MD live_VB V] ._. S]
```

M06 715

```
'_*' [S[Na we_PPlAS Na][V shall_MD soon_RB miss_VB V][N the_ATI trees_NNS
N][R no_RB longer_RBR R] ._. S]
```

N04 403

```
[S[N acute_JJ neurasthenia_NN N] ,_, [V said_VBD V] [N the_ATI surgeon-
rear-admiral_NN N] ._. S]
```

P09 1097

```
[S[Na I_PPlA Na][V guess_VB V][Fn[Na I_PPlA Na][V 'd_HVD better_RBR [VB&
go_VB [VB+ and_CC break_VB VB-]VB&]V] [N it_PP3 N] [P to_IN [N the_ATI
boys_NNS [P in_IN [N the_ATI lab_NN N]P] N]P]Fn] ._. S]
```

R06 413

```
[S[N the_ATI family_NN group_NN N][R then_RN R][V sat_VBD V][R down_RP R][P
for_IN [N a_AT late_JJ lunch_NN N]P][Fa before_CS [N the_ATI father-in-
law_NN N][V telephoned_VBD V][N the_ATI police_NNS N]Fa] ._. S]
```

the whole LOB Corpus. This is because the Parsed Corpus was produced by automatic syntactic analysis, followed by (repeated) manual correction, and the automatic syntactic analysis could not be performed on long sentences. However, apart from this restriction of length, the Parsed Corpus is broadly representative of British English.

The Lancaster Parsed Corpus is made available on diskette, accompanied by a Manual of Information giving details of the contents of the corpus, and the coding schemes used. With the Parsed Corpus itself, the Manual is distributed both in machine-readable form and in printed (hard copy) form. See further the order form accompanying the journal.

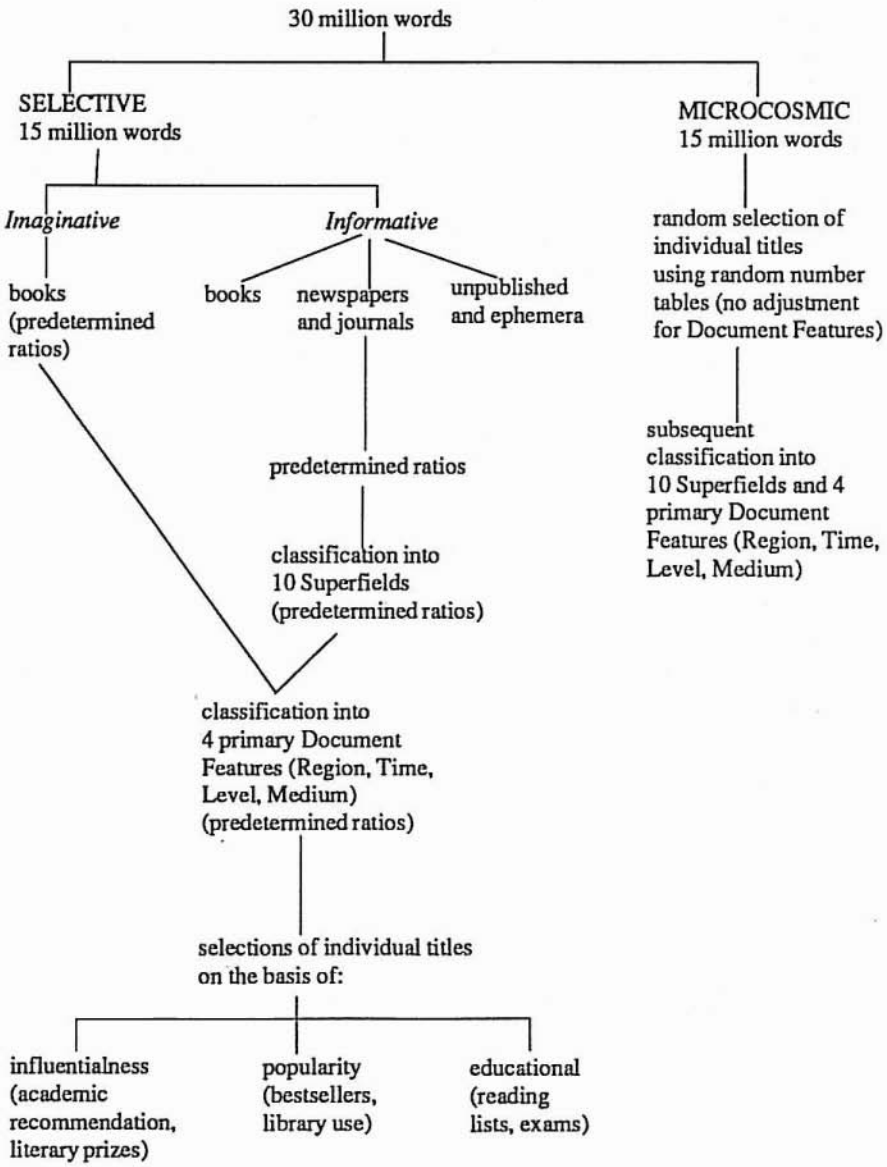
The Longman/Lancaster English Language Corpus and the Longman Corpus of Learners' English

Steve Crowdy
Longman

The Longman/Lancaster English Language Corpus contains 30 million words of twentieth-century English texts, covering American and British English predominantly, but also including other major varieties of native-speaker English. The composition of the Corpus has been determined by the Longman Corpus Committee, under the direction of Della Summers, with advice from a number of academics, notably Geoffrey Leech and Sir Randolph Quirk.

The Corpus is structured in two equal parts. The first is the *selective* corpus, with texts chosen on the basis of their occurrence on examination reading lists, set reading lists for the study of English, bestseller lists, library borrowing figures, and so on. Recommendations for specific titles have been gathered from academics and subject specialists in a variety of disciplines. In addition, there is text from a wide variety of periodicals as well as ephemeral material such as leaflets, letters, and packaging. The second half is the *microcosmic* corpus, which is collected by random sampling from a complete listing of books currently in print. The standard size of text taken from the sources is 40,000 words. Where the whole document is less than 40,000 words, the whole text is captured.

Longman/Lancaster English Language Corpus



The position of each 40,000-word block is varied at random so that there is a roughly equal number of beginning, middle, and end blocks.

The Corpus is being used by Longman for dictionaries and other books on language, and is also available more generally for academic research. Researchers wishing to purchase the Corpus should send the following details: a description of the intended research use; full names and academic details of researchers using the Corpus; details of source of research funding. Commercial use and onward transmission of the Corpus are not permitted.

The Longman Corpus of Learners' English is a large, computerized databank containing samples of written English produced by speakers of other languages. This Corpus currently stands at 2 million words, but we aim to build it up to 10 million in the near future. Contributions have been collected from over 70 countries worldwide at all levels from beginners to high proficiency. The Corpus provides objective information about those aspects of English grammar, lexis and usage which students find particularly difficult. Because the material is coded, it is possible to identify specific error trends that are characteristic of students at a particular level of competence or from a particular language background. It can therefore focus on a selected group, such as Intermediate Spanish students, or be used to analyse errors across the entire gamut of learners. The Corpus is available for academic research on the same terms as the Longman/Lancaster English Language Corpus.

For more information, contact Steve Crowdy or Della Summers at the following address:

Longman Dictionaries
Longman House
Burnt Mill Harlow
Essex CM20 2JE
England

The British National Corpus

The British National Corpus initiative is a major collaborative venture, the goal of which is the creation of a corpus of 100 million words of contemporary spoken and written British English. The project started in January 1991 and is to run for just over three years.

The participants in the project are Oxford University Press, which

leads the consortium, Longman Group UK Ltd, W & R Chambers, the British Library and the universities of Oxford (Oxford University Computing Service) and Lancaster (Unit for Computer Research on the English language). An Advisory Council under the chairmanship of Dr Anthony Kenny, President of the British Academy, oversees the project.

The texts for the British National Corpus are chosen to form a representative cross-section of a wide range of styles of current written and spoken English. Novels, magazines, technical manuals, ordinary conversation, lectures, advertisements, textbooks, radio and TV broadcasts are among the categories collected. The spoken material will include a demographic sample of everyday speech from British English speakers in the UK.

Once the 100 million words are collected, the Corpus will be analysed computationally and grammatical labels will be added to every word. To ensure that the Corpus will be usable as widely as possible, the texts will be stored and distributed according to the internationally recognised encoding guidelines being defined by the Text Encoding Initiative. A suite of corpus processing tools will be developed, which can be used for searching and retrieving information from the Corpus.

The Corpus will be used for work in lexicography and language technology and for language research in general. The intention is to make the finished Corpus available to anyone who requires a data sample of this type for their linguistic research, after a two-year period in which the material is reserved for the consortium members.

For further information, contact:

Jeremy Clear
Oxford University Press
Walton Street
Oxford OX2 6DP
Tel: +44 865 56767
Fax: +44 865 56646
E-mail: jhclear@vax.ox.ac.uk

The Cambridge Language Survey

Paul Procter
Cambridge University Press

The Cambridge Language Survey (CLS) is an international multilingual survey of language being organised by a growing consortium of publishers and industrial companies, the coordinating partner being Cambridge University Press (contact Paul Procter). The chairman of CLS is Sir John Lyons, and the UK partners include Ted Briscoe (Cambridge University), Sidney Greenbaum (International Corpus of English) and Reinhard Hartmann (University of Exeter). The overseas partners include Acquilex members: University of Amsterdam (Meijs, Vossen), Universitat Politecnica de Catalunya (Rodriguez, Verdejo), Universita di Pisa (Zampolli, Calzolari), Bibliograf, Barcelona (Accorda), Van Dale Lexicografie, Utrecht (Moerland). Links are being established with the Consortium for Lexical Research, Las Cruces, New Mexico (Wilks).

CLS is developing monolingual and multilingual dictionaries and software tools which will be available to the Natural Language Processing community at only nominal cost, subject to safeguards of copyright. The computer system being used is based on Novell networked IBM-compatible PCs, using a customisable relational database called Advanced Revelation.This system should be runnable by any scholar or institution with PCs. CLS is multilingual, treating all the languages involved (English, French, German, Spanish, Italian, Dutch, Japanese) with equal rigour.

Objectives

The objectives provide for:

PUBLISHING PARTNERS who are interested mainly in bilingual and multilingual publishing in book and electronic form, and in tools to assist the publishing process.

ACADEMIC PARTNERS who need data and systems for scholarly research, including multilingual language corpora, and tools for their analysis.

INDUSTRIAL PARTNERS who require processes to help to build products and services, and to exploit technological advances as they occur. Tools

of the survey include translation aids for technical manuals. Computer companies will obtain materials for incorporation into commercial software.

Areas of research

A central activity is the obtaining of reliable data on frequency of meaning using automated and manual sense-tagging, for all the languages, and coding both dictionaries and corpora. This involves concordancing and tagging software. The frequency of all types of collocation with particular meanings of particular word forms will be analysed statistically.

Core vocabulary

The steps taken will be:

1. within each language, to establish a core set of high-frequency meanings;
2. to cross-map equivalent meanings among the languages of the group;
3. to document degree of equivalence where one-to-one mappings are inappropriate.

This will result in appropriate semantic links on computer between items in different languages. Part of the results of these studies will be an adequate documentation of lexical interference (false friends).

Machine-tractable dictionaries in each language

These dictionaries are a primary tool for automating the process of assigning meanings to word forms in running text, based on a rich set of linguistic coding (see below).

Cultural data

We are collecting a body of material exemplifying cultural assumptions and allusions built into the various languages.

Language variety

CLS will look at differences of language variety between languages for the same semantic areas (e.g. differences of formality level).

Words in groups

'Words in groups' is adopted as a general term for lexical items consisting

of more than one word, and embraces different kinds of language including idioms, phrases, collocations, proverbs, quotations and allusions. The adequate documentation of these is a neglected area.

A universal linguistic coding

This is in development, and will be compatible with TEI and DEI standards, so that it becomes of general use.

1. Coding is hierarchical, using pop-down menus to select the appropriate level
2. There will be different coding systems for different parts of the lexicon.
3. The types of coding include:
   a) syntactic and grammatical (e.g. verb complementation)
   b) semantic with subcategories
      subject
      thesaurus (synonym etc. sets)
      selectional restrictions (e.g. subject / object)
      semantic relations
   c) morphological (e.g. inflections, combining potential of morphemes, part of speech components of compounds)
   d) orthographic (e.g. number of syllables, consonant clusters)
   e) phonologial / stress
   f) etymological (e.g. language of origin, cognates, language described)
   g) style
      national restriction
      regional
      level (e.g. formality)
      attitude (e.g. derogatory)
   h) status (e.g. neologism, proper name, cross-reference)
   i) collocation
   j) frequency

Products of the survey

analysed corpora in the various languages, parallel and aligned
machine-tractable dictionary databases
software tools, for a whole range of applications
electronic products for the whole range of current and future systems, including information sources (such as dictionaries and encyclopedias)

on CD-ROM, floppy disc and tape, CDI, hand-held solid state, hand-held disc-based, etc.
Applications of the research include:

EDUCATION
Education (user needs)
Language interference studies
Specialised lexicons
Language acquisition
Translation aids and systems

NATURAL LANGUAGE PROCESSING
Parsing (semantic and syntactic analysis)
Text understanding
Text generation
Discourse understanding
Machine-assisted translation leading to machine translation
Speech recognition and synthesis

For more information on the Cambridge Language Survey, contact:

Paul Procter,
Senior Editor,
International Dictionaries,
Cambridge University Press,
Edinburgh Building,
Shaftesbury Road,
Cambridge CB2 2RU,
England.

E-mail:psp10@cam.phx.ac.uk

The Bank of English

The Cobuild group at the University of Birmingham reports that they are going to raise the size of their corpus 'by an order of magnitude and move into the hundreds of millions of words' (circular letter from Professor John Sinclair, May 1991). The material will be used for research and development in grammar and lexicography. John Sinclair states that the material should be available for general use and invites

scholars 'to spend periods of study in a pleasant environment in the same building as Cobuild'.

The Georgetown University Catalogue of Projects in Electronic Text

The Center for Text & Technology at Georgetown University, under the direction of Dr. Michael Neuman, maintains an electronic catalogue of projects in electronic text in the humanities. The database includes a variety of information on the many collections of literary works, historical documents, and linguistic data sources which are available from commercial vendors and the scholarly community throughout the world. The database is written in Ingres and resides on a VAX 8700 computer at Georgetown University. The database may be searched by using Telnet or a modem. In addition, searches of the catalogue are performed on request, and updated lists of projects and addresses are posted regularly on the HUMANIST electronic bulletin board and distributed through surface and electronic mail.

For further information about the project please contact:

James A. Wilderotter II, Project Assistant
The Center for Text and Technology
Academic Computer Center
238 Reiss Science Building
Georgetown University
Washington, DC 20057
phone: (202) 687-6096
electronic mail: wilder@guvax.bitnet
wilder@guvax.georgetown.edu

The Center for Electronic Texts in the Humanities

Susan Hockey
Rutgers and Princeton Universities

The Center for Electronic Texts in the Humanities (CETH) was established in October 1991 by Rutgers and Princeton Universities with external support from the Andrew W Mellon Foundation and the National Endowment for the Humanities. Initial funding is for three years. Until now, developments in humanities computing and the compilation of resources and tools to support research in this area have been scattered, particularly in North America. The need to provide an on-going focus of interest for those who are involved in the creation, dissemination and use of electronic texts in the humanities has been recognized for some time. The Center is intended to serve that need. It will act as a national node on an international network of centers and projects which are actively involved in the handling of electronic texts.

Three major areas of activity are planned for the Center's start-up period:

(1) The inventory of machine-readable texts in the humanities which was begun by Marianne Gaunt at Rutgers in 1983. At present the inventory contains some 1,600 records which are held on RLIN. The Center is developing the inventory by reviewing the records which are already catalogued, some of which are several years old, as well as adding new ones. The records will be made available in other forms such as a file on Internet, a database and a printed publication. Priority will be given to cataloguing texts which are available commercially. Other sources of information such as *The Humanities Computing Yearbook*, journals, newsletters and electronic bulletin boards will be used to obtain information about texts.

Many of the records of the Oxford Text Archive have now been catalogued for the inventory, and the Center will also collaborate with the Center for Text and Technology at Georgetown University which has an online Catalogue of Projects in Electronic Text. Further substantial information is being obtained from the sections on corpora, text collections and individual texts within the survey of machine-readable texts organized by Antonio Zampolli and Donald Walker on behalf of the major text

analysis computing organizations.

(2) The acquisition and dissemination of text files. The Center will concentrate on a series of good quality texts which can be made available over Internet via suitable retrieval software and with appropriate copyright permissions. Our present plans are to begin with a collection of works of American Literature and History. The Center's texts will be encoded according to the Guidelines of the Text Encoding Initiative (TEI). By using only the TEI encoding format, the Center will be able to play a leading role in the testing, evaluation and dissemination of the Guidelines, particularly for humanities and literary material.

(3) Educational programs for humanities computing. The Center will also encourage more effective use of methodologies for research and instruction using electronic texts, by establishing a series of educational programs which are intended for faculty, librarians, computer staff, and graduate students. To begin with, a series of seminars on existing projects is being organized in the host universities which will highlight successful ways of introducing the computer into traditional humanities scholarship.

The Center is also planning summer seminars in humanities computing, the first of which will take place on 9-21 August 1992 at Princeton. This seminar is intended for those who have some basic computing experience, e.g. word processing and electronic mail but little or no experience of computers in a research environment. It will cover topics such as text encoding, methods of text acquisition, concordances, text retrieval, preparing critical editions and hypertext with practical work using software such as TACT and Micro-OCP. The seminar will also look at the current generation of software tools and then go on to examine what is needed to make these tools better for research applications in the humanities. The tutors will be Dr Willard McCarty of the Centre for Computing in the Humanities, University of Toronto, and Susan Hockey.

The provision of information services will also be an important activity for the Center, which will produce a regular newsletter in paper and electronic form as well as supporting an electronic bulletin board which will focus on the Center's own activities.

A major objective of the Center in the long term is to take a leading role in a partnership with centers and projects in Europe, Japan and elsewhere to conduct a feasibility study to establish ground rules for handling electronic texts, and then to establish mechanisms which can be used by all who have an interest in such material. We need to know how to preserve and document electronic texts and to maintain them whilst keeping up with new technological developments, also how to

deal with ownership issues both of texts which already exist and of future ones.

The operations of the Center are divided between Rutgers and Princeton Universities, with the administrative headquarters at Rutgers in the Alexander Library, and the computing operations mainly at Princeton. The Center also has an office in the Firestone Library at Princeton.

For further information please contact:

Susan Hockey
Director
Center for Electronic Texts in the Humanities,
169 College Ave, New Brunswick, NJ 08903
phone: (908) 932-1384
fax: (908) 932-1386
electronic mail: ceth@zodiac.rutgers.edu.

New Oxford Text Archive Catalogue

The new Catalogue is available in paper form by post from Oxford Text Archive, Oxford University Computing Services, 12 Banbury Rd, Oxford OX2 6NN. It is also available in electronic form, either as a formatted file for display at a terminal or in a tagged form using SGML. These files are now available from a number of different places as follows:

(1) on the Oxford VAX Cluster as
OX$DOC:TEXTARCHIVE.LIST and **OX$DOC:TEXTARCHIVE.SGML**

(2) on the Internet, these files are available for anonymous FTP from **black.ox.ac.uk** (129.67.1.165) (and elsewhere) in the directory **ota**. A number of other files are available from the same place.

(3) via ListServ, e.g. from
LISTSERV@BROWNVM and **FILESERV@HD.UIB.NO** which make the files available under the names **OTALIST FORMAT** and **OTALIST SGML**

(4) on JANET you can consult the list interactively on HUMBUL or download it from BUBL.

For more information, send a note to **ARCHIVE@VAX.OX.AC.UK**

News on the Text Encoding Initiative (TEI)

The new version of the guidelines for text encoding and interchange is due to be published in 1992. It will consist of three major parts: a prose specification, a reference section, and full document type definitions (DTDs). Apart from defining general mechanisms for text encoding and a set of core tags, the guidelines will propose base tags for prose, verse, drama, spoken text transcriptions, letters and memos, printed dictionaries, lexical data, terminological data, and language corpora and other collections. Additional topics handled by the guidelines are the encoding of linguistic and literary interpretation and analysis, text criticism and apparatus, etc. For news on the Text Encoding Initiative, subscribe to the discussion list TEI-L@UICVM.BITNET or get in touch with one of the TEI editors:

Michael Sperberg-McQueen
U35395@UICVM.BITNET

Lou Burnard
LOU@VAX.OX.AC.UK.

The CHILDES CD-ROM

The CHILDES Project (see the article by Brian MacWhinney, *ICAME Journal* 14 (1990), pp. 3-25) has released a CD-ROM in ISO 9660 format which can be read by Macintosh and MS-DOS machines which have a CD-ROM reader. The single disk contains the whole database (child language material), the programs, and the CHILDES/BIB system. The CHILDES CD-ROM is available free of charge.

For more information, contact:

Brian MacWhinney
Department of Psychology
Carnegie Mellon University
Pittsburgh, Pennsylvania 15213
USA

The ACL/DCI CD-ROM

The Data Collection Initiative of the Association for Computational Linguistics (see *ICAME Journal* 14, 1990, pp. 110-111) har released a CD-ROM with texts for use in linguistic research. The texts include *Wall Street Journal* material from 1987-89, the *Collins English Dictionary* (1979), and material from the Penn Treebank (cf. p. 112 in this issue). The CD-ROM can be obtained at a low charge from the Association for Computational Linguistics.

To obtain an order form and a copy of the user agreement, contact:

Rafi Khan
619 Williams Hall
University of Pennsylvania
Philadelphia, PA 19104-6305
USA

E-mail: khanr@unagi.cis.upenn.edu

The ICAME CD-ROM

Knut Hofland
Norwegian Computing Centre for the Humanities

The ICAME Collection of English Language Corpora is a new CD-ROM produced and distributed by the Norwegian Computing Centre for the Humanities. It includes the following corpora (for some information on these corpora, see pp. 141–143):

Brown Corpus: Bergen text version I and II, for MS-DOS, Macintosh and Unix. A modified Bergen version II indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

LOB Corpus: Tagged and untagged original text versions, for MS-DOS, Macintosh and Unix. A tagged horizontal version indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

Kolhapur Corpus: Text version for MS-DOS, Macintosh and Unix. A version indexed by WordCruncher 4.4 for MS-DOS.

London-Lund Corpus: Original text version for MS-DOS, Macintosh and Unix. An edited version indexed by WordCruncher 4.4 and TACT for MS-DOS and Free Text Browser for Macintosh.

Helsinki Corpus: Text version for MS-DOS, Macintosh and Unix. 1-file, 3-file and 11-file versions indexed by WordCruncher 4.4 and TACT for MS-DOS.

As the material is provided in a number of versions, it should be easy to use. The following programs are distributed with the disc: WordCruncher View, TACT, and Free Text Browser.

The disc contains a number of information files, including full lists of texts for the Brown, LOB, and Kolhapur corpora, and the list of speakers for the London-Lund Corpus. It also contains information on network resources, such as discussion lists and sites for downloading of programs, Netnews, lists of electronic text projects and some linguistic freeware programs. Manuals for the Helsinki and London-Lund corpora are distributed with the disc. See further the order form accompanying this journal.

ICAME services

The ICAME network server

Knut Hofland
Norwegian Computing Centre for the Humanities

The machine nora.hd.uib.no has been established as a mail-based server for the Norwegian Computing Centre for the Humanities. Information is grouped in different catalogues, some of which have information only in Norwegian. The relevant catalogues for ICAME are icame, ncch, info, pc, mac, and unix.

The server holds information about material available, some text samples, order forms, an ICAME bibliography, a survey of text corpora, programs and documentation, and network addresses.

The server is called FILESERV and runs the DECWRL archive server. FILESERV accepts three types of commands, and several commands can be placed in the body of the mail message. However, the results will be sent in one file, so do not request several large files in one message. The commands (can be sent in the Subject line or body):

Help	Help file
Index	Top level index
Index <catalogue>	Index for a catalogue
send <catalogue> <filename>	Fetch a file in a catalogue

Example: We want to get the files icame.cond and icame.material in the catalogue icame. Send the following note:

To: fileserv@hd.uib.no
Subject: whatever (or a command)

send icame icame.cond
send icame icame.material

The files are also available via anonymous FTP from nora.hd.uib.no (129.177.24.42).

Texts available through ICAME

The following corpora are currently available through the International Computer Archive of Modern English (ICAME). For information on the CD-ROM, see further p. 139.

Brown Corpus, untagged text format I (available on tape, diskette, and CD-ROM): A revised version of the Brown Corpus with upper- and lower-case letters and other features which reduce the need for special codes and make the material more easily readable. A number of errors found during the tagging of the corpus have been corrected. Typographical information is preserved; the same line division is used as in the original version from Brown University except that words at the end of the line are never divided.

Brown Corpus, untagged text format II (tape, diskette, and CD-ROM): This version is identical to text format I, but typographical information is reduced and the line division is new.

Brown Corpus, KWIC concordance (tape and microfiche): A complete concordance for all the words in the corpus, including word statistics showing the distribution in text samples and genre categories. The microfiche set includes the complete text of the corpus.

Brown Corpus, other versions (diskette and CD-ROM): See p. 139. The WordCruncher version is described in an article by Randall Jones, *ICAME Journal* 11, pp. 44-47.

LOB Corpus, untagged version, text (tape, diskette, and CD-ROM): The LOB Corpus is a British English counterpart of the Brown Corpus. It contains approximately a million words of printed text (500 text samples of about 2,000 words). The text of the LOB Corpus is not available on microfiche.

LOB Corpus, untagged version, KWIC concordance (tape and microfiche): A complete concordance for all the words in the corpus. It includes word statistics for both the LOB Corpus and the Brown Corpus, showing the distribution in text samples and genre categories for both corpora.

LOB Corpus, tagged version, horizontal format (tape, diskette, and CD-ROM): A running text where each word is followed immediately by a word-class tag (number of different tags: 134).

LOB Corpus, tagged version, vertical format (tape and CD-ROM):

Each word is on a separate line, together with its tag, a reference number, and some additional information (indicating whether the word is part of a heading, a naming expression, a quotation, etc).

LOB Corpus, tagged version, KWIC concordance (tape and microfiche): A complete concordance for all the words in the corpus, sorted by key word and tag. At the beginning of each graphic word there is a frequency survey giving the following information: (1) total frequency of each tag found with the word, (2) relative frequency of each tag, and (3) absolute and relative frequencies of each tag in the individual text categories.

LOB Corpus, other versions (diskette and CD-ROM): See p. 139.

Lancaster Parsed Corpus (tape and diskette): See the description on p. 124.

London-Lund Corpus, complete text (computer tape, diskette, and CD-ROM): The London-Lund Corpus contains samples of educated spoken British English, in orthographic transcription with detailed prosodic marking. It consists of 100 'texts', each of some 5,000 running words. The text categories represented are spontaneous conversation, spontaneous commentary, spontaneous and prepared oration, etc. The original version of the London-Lund Corpus (87 texts) is no longer available. As regards the versions available, see p. 139.

London-Lund Corpus, KWIC concordance I (computer tape): A complete concordance for the 34 texts representing spontaneous, surreptitiously recorded conversation (text categories 1-3), made available both in computerized and printed form (J. Svartvik and R. Quirk (eds.) *A Corpus of English Conversation*, Lund Studies in English 56, Lund: C.W.K. Gleerup, 1980).

London-Lund Corpus, KWIC concordance II (computer tape): A complete concordance for the remaining 53 texts of the original London-Lund Corpus (text categories 4-12).

London-Lund Corpus, supplement (tape and diskette): The 13 texts not included in the original version of the London-Lund Corpus. See the presentation by Sidney Greenbaum, *ICAME Journal* 14 (1990) pp. 108-110.

Melbourne-Surrey Corpus (tape or diskette): 100,000 words of Australian newspaper texts (see the article by Ahmad and Corbett, *ICAME Journal* 11, pp. 39-43).

Kolhapur Corpus, original version (tape, diskette, and CD-ROM): A million-word corpus of printed Indian English texts. See the article by

S.V. Shastri, *ICAME Journal* 12, pp. 15-26.

Kolhapur Corpus, other versions (diskette and CD-ROM): See p. 139.

Lancaster/IBM Spoken English Corpus (tape or diskette): A corpus of approximately 52,000 words of contemporary spoken British English. The material is available in orthographic and prosodic transcription and in two versions with grammatical tagging (like those for the LOB Corpus). There is an accompanying manual. See further *ICAME Journal* 12, pp. 76-77.

Polytechnic of Wales Corpus (tape or diskette): Orthographic transcriptions of some 61,000 words of child language data. The corpus is parsed according to Hallidayan systemic-functional grammar. There is no prosodic information. See further *ICAME Journal* 13 (1989), p. 20ff, and 15 (1991), pp. 55-62.

Helsinki Corpus (tape, diskette, and CD-ROM): A selection of texts covering the Old, Middle, and Early Modern English periods, totalling 1.5 million words. See the article by Merja Kytö and Matti Rissanen on pp. 7-27. As regards the versions available, see p. 139.

Most of the material has been described in greater detail in previous issues of our journal. Prices and technical specifications are given on the order forms which accompany the journal. *Note that tagged versions of the Brown Corpus cannot be obtained through ICAME. The same applies to audio tapes for the London-Lund Corpus, the Lancaster/IBM Spoken English Corpus, and the Polytechnic of Wales Corpus.*

There are available printed manuals for the LOB Corpus (the original manual and a supplementary manual for the tagged version), the Helsinki Corpus, and the London-Lund Corpus. Printed manuals for the Brown Corpus cannot be obtained from Bergen. Users of the London-Lund material are also recommended to consult J. Svartvik (ed.). *The London-Lund Corpus: Description and Research*, Lund University Press, 1990.

A manual for the Kolhapur Corpus can be ordered from: S.V. Shastri, Department of English, Shivaji University, Vidyanagar, Kolhapur-416006, India. The price of this manual is US $15 (including airmail charges). Payment should be sent along with the order by cheque or international postal order drawn in favour of The Registrar, Shivaji University, Kolhapur.

Conditions on the use of ICAME corpus material

The following conditions govern the use of corpus material distributed through ICAME:

1. No copies of corpora, or parts of corpora, are to be distributed under any circumstances without the written permission of ICAME.
2. Print-outs of corpora, or parts thereof, are to be used for bona fide research of a non-profit nature. Holders of copies of corpora may not reproduce any texts, or parts of texts, for any purpose other than scholarly research without getting the written permission of the individual copyright holders, as listed in the manual or record sheet accompanying the corpus in question. (For material where there is no known copyright holder, the person(s) who originally prepared the material in computerized form will be regarded as the copyright holder(s).)
3. Commercial publishers and other non-academic organizations wishing to make use of part or all of a corpus or a print-out thereof must obtain permission from all the individual copyright holders involved.
4. Publications making use of the material should include a reference to the relevant corpus (or corpora), giving the name of the corpus and the distributor.

Information for contributors

Language. All contributions should be in English. Contributors whose native language is not English should have their manuscripts gone through by a native speaker before submission.

Format. Contributions should preferably be submitted as ASCII files on diskette, together with a printout made from your word-processing system. As regards other possible formats, consult the editors before submission of your manuscript.

Headings. The title of the paper should be followed by the author's name and academic affiliation. Sections and sub-sections should be

numbered. Headings should **not** be singled out typographically (by boldface, capitalization, or the like).

Tables and figures should be numbered and titled. They should always be referred to by their number, **not** by expressions like 'see the diagram below' or 'in the following table'. Tables should be submitted in a separate file. Drawings, graphs, and other illustrations must be reproducible originals.

Quotations. Use single quotation marks, except for quotes within quotes. Long quotations should be indented and given without quotation marks.

Examples should normally be numbered and set apart from the text following standard linguistic practice. Short examples in the running text (words or phrases) should be underlined.

Notes should be placed at the end of the paper. References to notes in the text should be indicated as follows: *1, *2, etc.

References should conform to standard linguistic practice. References in the text should follow this pattern: Francis (1979: 110) defines a corpus as 'a collection of texts assumed to be representative of a given language, dialect, or other subset of a language, to be used for linguistic analysis'. The list of references at the end of the paper should be presented as shown by these examples:

Altenberg, Bengt. 1984. Causal linking in spoken and written English. *Studia Linguistica* 38:20-69.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Renouf, Antoinette. 1987. Corpus development. In *Looking up: An account of the COBUILD Project in lexical computing*, ed. by J. M. Sinclair. 1-40. London & Glasgow: Collins ELT.

Tottie, Gunnel, and Ingegerd Bäcklund (eds.). 1986. *English in speech and writing: A symposium*. Studia Anglistica Upsaliensia 60. Stockholm: Almqvist & Wiksell.

Authors should be given with their full first names, unless they always use the initials themselves.

Reviews. The heading of a review should contain the information shown in the following example:

Roger Garside, Geoffrey Leech, and Geoffrey Sampson (eds.). *The computational analysis of English: A corpus-based approach*. London:

Longman, 1987. 196 pp. ISBN 0-582-29149-6. Reviewed by Gunnel Källgren, University of Stockholm.

Review articles should have a title, followed by the author's name and affiliation, and the information on the book(s) reviewed, as shown above.

Submission, books for review. Contributions, as well as books for review, should be sent to one of the editors:

Stig Johansson
Department of British
and American Studies
University of Oslo
P.O. Box 1003
Blindern
N-0315 Oslo 3
Norway
E-mail: stigj@hedda.uio.no

Anna-Brita Stenström
Department of English
University of Bergen
Sydnesplass 9
N-5007 Bergen
Norway
stenstroem@hf.uib.no

The editors are grateful for any information or documentation which is relevant to the field of concern of ICAME.

ICAME Journal is published by the Norwegian Computing Centre
for the Humanities (Humanistisk datasenter)
Address: Harald Hårfagres gate 31, N-5007 Bergen, Norway.
Telephone: +47 5 212954 Telefax: +47 5 322656
E-mail: icame@hd.uib.no
ISSN 0801-5775